Geometry-Adaptive Explainer for Faithful Dictionary-Based Interpretability under Distribution Shift

Sungjun Lim[a] **Heedong Kim**[a] **Andrew Lee**[b,*] **Kyungwoo Song**[a,*]

[a]Yonsei University [b]Harvard University

Abstract

Mechanistic interpretability aims to explain a model's behavior by identifying causally responsible internal structures. Dictionary-based explainers such as sparse autoencoders and transcoders are a primary tool, but their faithfulness under out-of-distribution (OOD) shift has received little systematic attention. We show that distribution shift rotates the subspace that the model actively uses, misaligning the explainer's dictionary trained on in-distribution (ID) activations. We formalize this misalignment as the **faithfulness gap**, a geometric distance between the ID dictionary and the OOD-active subspace, and show that it bounds excess OOD projection loss. To reduce this gap, we propose the **Geometry-Adaptive Explainer (GAE)**, which realigns the dictionary with the OOD-active subspace while preserving the original sparse feature activations. This requires only unlabeled OOD activations and no gradient updates. For the adapted dictionary, we also bound the excess projection loss caused by any remaining faithfulness gap. Empirically, GAE even matches or surpasses all training-based baselines in causal faithfulness across multiple models and OOD settings.[1]

1 Introduction

Mechanistic interpretability aims to explain a model's behavior by identifying internal structures that are causally responsible for its outputs [1–3]. A primary approach is to train an explainer, a post-hoc module such as a sparse autoencoder (SAE) [4] or transcoder [5], that decomposes hidden activations into sparse combinations of learned feature directions (a dictionary). These dictionary-based explainers have recently been scaled to large language models [6, 7] and used to uncover interpretable feature circuits [8]. A central requirement for such explanations is faithfulness: they should accurately reflect the computations the model actually uses [9, 1].

When a model encounters out-of-distribution (OOD) inputs, the dictionary learned in-distribution (ID) can no longer capture the directions the model actively uses [10]. Prior work has addressed this vulnerability from two angles, each limited in scope. Attribution-level robustness studies [11–14] focus on input perturbations rather than hidden-state geometry. Dictionary-explainer remedies such as retraining on the model's own generations [15], upweighting rare concepts [16, 17], and adding residual modules [18] remain heuristic, without diagnosing the underlying misalignment. Consequently, the mechanism driving this failure and a principled correction remain unaddressed.

In this work, we identify a geometric mechanism for OOD faithfulness degradation in dictionary-based explainers. These explainers learn their feature directions from ID activations, so their dictionaries reflect the geometric structure of ID hidden representations [19]. This structure is captured by

*Corresponding authors: andrewlee@g.harvard.edu, kyungwoo.song@gmail.com.

[1]Code is available at: https://github.com/MLAI-Yonsei/GAE/

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

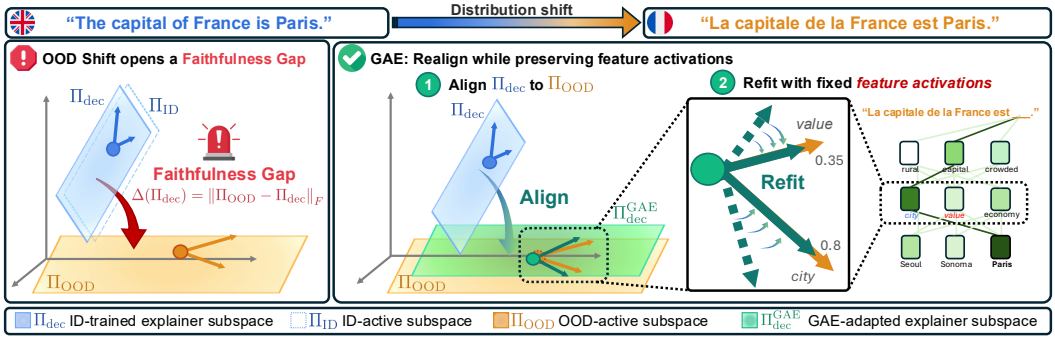


Figure 1: **Faithfulness gap and GAE.** *Left*: distribution shift (illustrated as a language change) rotates the OOD-active subspace Π_{OOD} away from the ID-trained explainer subspace $\Pi_{\mathrm{dec}} \approx \Pi_{\mathrm{ID}}$, opening a faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$. *Right*: GAE uses two steps. Step 1 rotates Π_{dec} onto the estimated OOD-active subspace via orthogonal Procrustes. Step 2 refits feature directions under an alignment penalty to match OOD activations while preserving the original sparse feature activations.

the second moment of hidden activations, which distribution shift typically alters [20], leaving the ID-trained dictionary misaligned with the OOD-active subspace. As illustrated in Figure 1 (left), we call this misalignment the **faithfulness gap**, a geometric distance between the ID dictionary and the OOD-active subspace. We prove that this gap bounds excess OOD projection loss and that it is itself upper-bounded by the magnitude of the second-moment shift when the explainer is aligned with the ID-active subspace. Reducing this gap tightens the projection-loss bound, motivating methods that directly realign the dictionary.

We instantiate this idea with the **Geometry-Adaptive Explainer (GAE)**, a closed-form, post-hoc method that targets the faithfulness gap using only unlabeled OOD activations. GAE first rotates the ID-trained dictionary so that its subspace aligns with the estimated OOD-active subspace, choosing the rotation closest to the original dictionary while preserving the original sparse feature activations (Step 1 in Figure 1). Consequently, the same input activates the same features with the same strengths before and after adaptation. A decoder refit then adjusts individual feature directions to match OOD activations while penalizing deviations from this alignment (Step 2 in Figure 1). The entire pipeline requires no gradient computation. We show that GAE's projection loss remains close to that of the best subspace of the same dimension when subspace estimation is accurate and decoder refitting introduces little additional misalignment (Corollary 1). Empirically, GAE matches or surpasses all training-based baselines in causal faithfulness across multiple language models and OOD settings, including full OOD retraining.

Our main contributions are summarized as follows:

- We identify the **faithfulness gap** as a geometric mechanism for OOD faithfulness degradation and show how it controls excess OOD projection loss.
- We propose **GAE**, a closed-form dictionary realignment method that targets the faithfulness gap while preserving the original sparse feature activations.
- Without any gradient computation, GAE matches or surpasses all training-based baselines, including full OOD retraining, in causal faithfulness across multiple models and diverse OOD settings.

2 Related Work

2.1 Faithfulness in Mechanistic Interpretability

Dictionary-based explainers such as SAEs [4, 21] and transcoders [5, 22] decompose hidden activations into sparse feature directions. Faithfulness is evaluated through feature ablation [9, 23, 24] and supports reliable circuit discovery and model editing [8, 25, 26]. Recent empirical reports show that SAE-based features underperform dense baselines on OOD downstream tasks [10].

2.2 Interpretability under Distribution Shift

Prior work studies the fragility of saliency maps under input perturbations [11, 12] and explanation consistency under shift [14, 13]. These analyses focus on attribution methods rather than hidden-space geometry. OOD detection reveals changes in hidden activations [20], which representation comparison methods can quantify [27–29]. The link to explainer faithfulness remains unexplored.

For dictionary-based explainers, proposed remedies include retraining on OOD data, training on the model's own generations [15], upweighting tail samples [16, 17], or adding residual capacity [18].

Subspace Alignment (SA) [30] and CORrelation ALignment (CORAL) [31] align source and target subspaces or covariance statistics for domain adaptation. GAE adapts a dictionary decoder for attribution and intervention while keeping the model and sparse encoder frozen.

3 Hidden-Space Geometry Shift and Faithfulness Degradation

When a neural network encounters OOD inputs, do its mechanistic explanations remain faithful? We show that they generally do not. Distribution shift alters the geometry of hidden activations, creating a misalignment between the directions the model actively uses and those the explainer was trained to reconstruct. We call this misalignment the faithfulness gap, and prove that **(i)** it grows at most proportionally with the second-moment shift for the ID-active subspace (Proposition 1), and **(ii)** it controls the reducible part of OOD projection loss (Proposition 2).

3.1 Setup and Faithfulness Gap

Target model and explainer. For an input X, the fixed pretrained target model produces the reconstruction target $h(X) \in \mathbb{R}^d$ at a designated layer. An encoder produces sparse feature activations $z(X) \in \mathbb{R}^k$, and a decoder (dictionary) $W_{\text{dec}} \in \mathbb{R}^{d\times k}$ reconstructs $h(X)$. SAEs encode $h(X)$ itself, while transcoders encode the MLP input to reconstruct its output $h(X)$; both share the form

$$\hat{h}(X) = W_{\text{dec}}\, z(X) + b_{\text{dec}}. \tag{1}$$

The encoder applies a sparsifying nonlinearity (e.g., ReLU or TopK) [21].

In-distribution, out-of-distribution, and second-moment shift. Let P_{ID} and P_{OOD} denote the ID training and OOD deployment distributions, respectively.

We characterize the shift via second-moment matrix $M_e := \mathbb{E}_{X\sim P_e}[h(X)h(X)^\top]$, $e \in \{\text{ID}, \text{OOD}\}$. The leading eigenvectors of M_e identify the directions that capture the most activation energy. Changes in these directions can therefore create a mismatch with the ID-trained dictionary. A second-moment shift occurs when $M_{\text{OOD}} \neq M_{\text{ID}}$, with magnitude $\|M_{\text{OOD}} - M_{\text{ID}}\|_F$.

Active subspace and faithfulness gap. Hidden activations concentrate energy along a few directions [32–34]. We write $\Pi_e = U_e U_e^\top$, where $U_e \in \mathbb{R}^{d\times r}$ contains the top-r eigenvectors of M_e, and call this the active subspace in environment e. Similarly, $\Pi_{\text{dec}} = U_{\text{dec}} U_{\text{dec}}^\top$, where $U_{\text{dec}} \in \mathbb{R}^{d\times r}$ contains the top-r left singular vectors of W_{dec}, is the explainer subspace.

Under second-moment shift, Π_{OOD} may diverge from Π_{ID}, opening a gap between the explainer subspace and OOD-active subspace. We call this misalignment the faithfulness gap, which controls excess OOD projection loss, as we formalize in Proposition 2:

Definition 1 (Faithfulness Gap)**.** *The faithfulness gap of an explainer subspace* Π_{dec} *under OOD is*

$$\Delta(\Pi_{\text{dec}}) := \|\Pi_{\text{OOD}} - \Pi_{\text{dec}}\|_F.$$

For $\Pi_{\text{dec}} = \Pi_{\text{ID}}$, second-moment shift upper-bounds this gap (Proposition 1).

3.2 Second-Moment Shift Bounds the Faithfulness Gap

The previous subsection defined the faithfulness gap as a geometric quantity. We now show that second-moment shift upper-bounds this gap for the ID-active subspace. In practice, explainers are trained on ID activations and deployed without modification [35, 36]. When an ID-trained explainer satisfies $\Pi_{\text{dec}} \approx \Pi_{\text{ID}}$ (empirically assessed in Appendix B.2, Table 6), its OOD faithfulness depends on how far Π_{ID} lies from Π_{OOD}. The following result bounds this distance.

Proposition 1 (Second-Moment Shift Bounds the Faithfulness Gap of the ID Explainer)**.** *Suppose that* M_{ID} *has eigengap* $\gamma_{\text{ID}} = \lambda_r(M_{\text{ID}}) - \lambda_{r+1}(M_{\text{ID}}) > 0$ *at rank* r*. Then*

$$\Delta(\Pi_{\text{ID}}) = \|\Pi_{\text{OOD}} - \Pi_{\text{ID}}\|_F \leq \frac{2}{\gamma_{\text{ID}}} \|M_{\text{OOD}} - M_{\text{ID}}\|_F.$$

The faithfulness gap of the ID-active subspace grows at most proportionally with the second-moment shift $\|M_{\text{OOD}} - M_{\text{ID}}\|_F$, with sensitivity controlled by the inverse eigengap $1/\gamma_{\text{ID}}$ (proof in Appendix A.1; empirical verification in Appendix B.4). Larger shifts permit a larger upper bound on the

faithfulness gap. For an ID dictionary whose projector differs from Π_{ID}, the triangle inequality gives $\Delta(\Pi_{\mathrm{dec}}) \leq \Delta(\Pi_{\mathrm{ID}}) + \|\Pi_{\mathrm{dec}} - \Pi_{\mathrm{ID}}\|_F$. We next relate this gap to excess projection loss.

3.3 The Faithfulness Gap Controls OOD Degradation

We now formalize the connection between the faithfulness gap and OOD faithfulness loss, showing that $\Delta(\Pi_{\mathrm{dec}})$ is the central quantity an adaptation method should target.

OOD faithfulness objective. An ideal explainer subspace Π_{dec} minimizes projection-level reconstruction error on OOD activations. We formalize this objective as

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) := \mathbb{E}_{X \sim P_{\mathrm{OOD}}} \|h(X) - \Pi_{\mathrm{dec}} h(X)\|_2^2. \tag{2}$$

This measures how much OOD activation is lost when projected onto Π_{dec}. $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$ is a projection-level surrogate for OOD faithfulness [23, 24].

Decomposition of $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$. To isolate the part of the OOD loss that the explainer can reduce, we decompose $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$ into two terms. Let $\mathcal{C}_r$ denote the set of rank-r orthogonal projectors in $\mathbb{R}^d$. For any $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$,

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) = \underbrace{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})}_{\text{irreducible}} + \underbrace{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})}_{\text{explainer-dependent}}. \tag{3}$$

The explainer-dependent component is nonnegative (proof in Appendix A.2), so Π_{OOD} minimizes $\mathcal{L}_{\mathrm{OOD}}$ over $\mathcal{C}_r$. The irreducible component depends on the target model and the OOD distribution, both fixed at deployment; adapting the explainer cannot reduce it. *The explainer-dependent component is the only part the explainer can reduce.*

Faithfulness gap as a tight proxy. The faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$ measures the distance between two subspaces, making it a direct optimization target. The next proposition shows that controlling $\Delta(\Pi_{\mathrm{dec}})$ is equivalent to controlling the explainer-dependent component.

Proposition 2 (Faithfulness Gap Controls the Explainer-Dependent Term)**.** *Assume that the OOD eigengap at rank r, $\gamma_{\mathrm{OOD}} = \lambda_r(M_{\mathrm{OOD}}) - \lambda_{r+1}(M_{\mathrm{OOD}}) > 0$. Then for any $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$,*

$$\frac{\gamma_{\mathrm{OOD}}}{2} \Delta(\Pi_{\mathrm{dec}})^2 \leq \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \leq \frac{\lambda_1(M_{\mathrm{OOD}}) - \lambda_d(M_{\mathrm{OOD}})}{2} \Delta(\Pi_{\mathrm{dec}})^2.$$

The lower bound shows that any nonzero gap incurs a positive cost (misalignment cannot be free); the upper bound limits this projection-loss cost quadratically in the gap. Together, they establish that controlling $\Delta(\Pi_{\mathrm{dec}})$ is equivalent to controlling the explainer-dependent projection loss (proof in Appendix A.3; empirical verification in Appendix B.3). *Reducing $\Delta(\Pi_{\mathrm{dec}})$ tightens an upper bound on excess projection loss, while improvements in nComp and nAOPC are evaluated empirically.* Section 4 introduces a method that directly targets this quantity.

4 Geometry-Adaptive Explainer (GAE)

Section 3 bounds excess OOD projection loss using the faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$. We now propose the Geometry-Adaptive Explainer (GAE), which targets $\Delta(\Pi_{\mathrm{dec}})$ by realigning the explainer's dictionary with the OOD-active subspace while preserving the original sparse feature activations. We present GAE's objective, a closed-form adaptation procedure, and a projection-loss bound.

4.1 Problem Formulation

Section 3.2 bounds the distance between the ID-active and OOD-active subspaces. Minimizing $\Delta(\Pi_{\mathrm{dec}})$ amounts to choosing a W_{dec} whose induced subspace aligns with Π_{OOD}. To restore faithfulness, we adapt the existing dictionary $W_{\mathrm{dec}}^{\mathrm{ID}} \in \mathbb{R}^{d \times k}$ using unlabeled OOD pairs $\{(h_i, z_i)\}_{i=1}^N$, estimating the OOD-active subspace $\widehat{\Pi}_{\mathrm{OOD}}$ from h_i. We seek W_{dec} that closes this gap while preserving the original sparse feature activations:

$$\min_{W_{\mathrm{dec}}} \underbrace{\mathcal{L}_{\mathrm{recon}}}_{\text{reconstruction}} + \lambda_{\mathrm{geom}} \underbrace{\|\widehat{\Pi}_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F^2}_{\text{subspace alignment}} + \lambda_{\mathrm{pres}} \underbrace{\|W_{\mathrm{dec}} - W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2}_{\text{dictionary preservation}}, \tag{4}$$

where $\mathcal{L}_{\text{recon}} = \frac{1}{N}\sum_{i=1}^{N} \|h_i - \hat{h}_i\|_2^2$ is the mean reconstruction error.[2] GAE preserves the original sparse feature activations by keeping the encoder fixed. For the same input, feature indices, activation values, and sparsity patterns remain unchanged [4]. The three terms serve complementary roles. The *reconstruction* term fits individual OOD activations, since subspace alignment alone does not guarantee sample-level reconstruction. The *subspace alignment* term directly targets the faithfulness gap $\Delta(\Pi_{\text{dec}})$, which Proposition 2 showed controls the explainer-dependent component. Dictionary preservation keeps the dictionary close to the original. Preserving feature activations does not ensure unchanged semantics or causal effects, which depend on decoder directions.

4.2 Dictionary Adaptation

Equation (4) is non-convex in W_{dec}, as Π_{dec} depends on it through a top-r SVD. GAE uses two closed-form steps (Algorithm 1). Step 1 enforces *subspace alignment* by rotating the ID-trained dictionary onto $\widehat{\Pi}_{\text{OOD}}$, choosing the rotation closest to the original. Step 2 refits the decoder via ridge regression, solving the *reconstruction* and *dictionary preservation* terms while penalizing deviations from this alignment.

Algorithm 1 GAE

Require: $W_{\text{dec}}^{\text{ID}} \in \mathbb{R}^{d\times k}$; OOD pairs $\{(h_i, z_i)\}_{i=1}^{N}$; λ_{geom}, λ_{pres}
Ensure: $(W_{\text{dec}}^{\text{GAE}}, b_{\text{dec}}^{\text{GAE}})$

$\widehat{M}_{\text{OOD}} \leftarrow \frac{1}{N}\sum_{i=1}^{N} h_i h_i^\top$
$U_{\text{dec}} \leftarrow$ top-r left singular vectors of $W_{\text{dec}}^{\text{ID}}$; $U_{\text{OOD}}^{(:r)} \leftarrow$ top-r eigenvectors of $\widehat{M}_{\text{OOD}}$
$G \leftarrow U_{\text{dec}}^\top W_{\text{dec}}^{\text{ID}} (W_{\text{dec}}^{\text{ID}})^\top U_{\text{OOD}}^{(:r)}$; $T^\star \leftarrow \widetilde{V}\widetilde{U}^\top$ from $\text{SVD}(G)$
$\widetilde{W}_{\text{dec}} \leftarrow U_{\text{OOD}}^{(:r)} T^\star U_{\text{dec}}^\top W_{\text{dec}}^{\text{ID}}$ ▷ Step 1: Subspace rotation
$W_{\text{dec}}^{\text{GAE}} \leftarrow$ Equation (8); $b_{\text{dec}}^{\text{GAE}} \leftarrow$ Equation (9) ▷ Step 2: Decoder refit

Step 1: Subspace rotation. Let $U_{\text{dec}} \in \mathbb{R}^{d\times r}$ be the top-r left singular vectors of $W_{\text{dec}}^{\text{ID}}$ (the explainer subspace defined in Section 3.1), and let $U_{\text{OOD}}^{(:r)}$ be the top-r eigenvectors of the empirical second-moment matrix $\widehat{M}_{\text{OOD}} = \frac{1}{N}\sum_{i=1}^{N} h_i h_i^\top$. We constrain the rotated dictionary to the form

$$\widetilde{W}_{\text{dec}}(T) \;=\; U_{\text{OOD}}^{(:r)}\, T\, U_{\text{dec}}^\top\, W_{\text{dec}}^{\text{ID}}, \qquad T \in \mathcal{O}_r, \tag{5}$$

where $\mathcal{O}_r = \{T \in \mathbb{R}^{r\times r} : T^\top T = I\}$. Every column of $\widetilde{W}_{\text{dec}}(T)$ lies in $\text{span}(U_{\text{OOD}}^{(:r)})$, so the column space of $\widetilde{W}_{\text{dec}}(T)$ is contained in this r-dimensional subspace. If $\text{rank}(W_{\text{dec}}^{\text{ID}}) \geq r$, $\widetilde{W}_{\text{dec}}(T)$ has rank r and its induced explainer subspace equals $\widehat{\Pi}_{\text{OOD}}$ exactly for any $T \in \mathcal{O}_r$. Among these rotations, we choose the one that keeps the dictionary closest to the original:

$$T^\star \;=\; \arg\min_{T\in\mathcal{O}_r} \left\|\widetilde{W}_{\text{dec}}(T) - W_{\text{dec}}^{\text{ID}}\right\|_F^2. \tag{6}$$

This is an orthogonal Procrustes problem [38]. Let $G = U_{\text{dec}}^\top W_{\text{dec}}^{\text{ID}} (W_{\text{dec}}^{\text{ID}})^\top U_{\text{OOD}}^{(:r)} \in \mathbb{R}^{r\times r}$, with SVD $G = \widetilde{U}\Sigma\widetilde{V}^\top$. Then $T^\star = \widetilde{V}\widetilde{U}^\top$ (derivation in Appendix A.4). Since the Step 1 dictionary has projector $\widehat{\Pi}_{\text{OOD}}$ by construction, the residual faithfulness gap reduces to the eigenspace estimation error. This yields a quantitative improvement over the unadapted ID explainer.

Theorem 1 (Improvement over ID Explainer)**.** *Suppose* $\gamma_{\text{OOD}} := \lambda_r(M_{\text{OOD}}) - \lambda_{r+1}(M_{\text{OOD}}) > 0$, $\Delta(\Pi_{\text{ID}}) > 0$*, and* $\widehat{\Pi}_{\text{OOD}} = \Pi_{\text{OOD}}$*. Then*

$$\mathcal{L}_{\text{OOD}}(\widehat{\Pi}_{\text{OOD}}) \;\leq\; \mathcal{L}_{\text{OOD}}(\Pi_{\text{ID}}) \;-\; \frac{\gamma_{\text{OOD}}}{2}\,\Delta(\Pi_{\text{ID}})^2. \tag{7}$$

This guarantees a Step 1 projection-loss improvement over the ID-active projector under exact subspace estimation. The proof is given in Appendix A.6.

[2] The full training objective for the dictionary-based explainer in Equation (1) includes a sparsity penalty on $\mathbf{z}_i$. GAE holds the encoder fixed, so $\mathbf{z}_i$ is constant with respect to W_{dec} and the sparsity term drops out, leaving only the reconstruction term. This also means GAE applies regardless of sparsity mechanism (ℓ_1, Top-K [7], JumpReLU [37], etc.).

Step 2: Decoder refit. Step 1 aligns the subspace but does not optimize sample-level reconstruction. Step 2 refits $\mathcal{L}_{\text{recon}}$ while penalizing deviations from the Step 1 alignment. With fixed feature activations z_i, reconstruction reduces to a linear least-squares problem in W_{dec} and b_{dec}. To keep the decoder geometrically aligned, we penalize decoder mass outside $\widehat{\Pi}_{\text{OOD}}$ with $\lambda_{\text{geom}}\|(I - \widehat{\Pi}_{\text{OOD}})\, W_{\text{dec}}\|_F^2$. To limit displacement from the aligned dictionary, the dictionary preservation penalty regularizes the decoder toward $\widetilde{W}_{\text{dec}}(T^\star)$ with $\lambda_{\text{pres}}\|W_{\text{dec}} - \widetilde{W}_{\text{dec}}(T^\star)\|_F^2$. The combined objective is convex and quadratic, yielding the closed-form solution

$$W_{\text{dec}}^{\text{GAE}} \;=\; \widehat{\Pi}_{\text{OOD}}\, C\, B^{-1} \;+\; (I - \widehat{\Pi}_{\text{OOD}})\, C\, (B + \lambda_{\text{geom}}\, I)^{-1}, \tag{8}$$

$$b_{\text{dec}}^{\text{GAE}} \;=\; \tfrac{1}{N}\textstyle\sum_i h_i \;-\; W_{\text{dec}}^{\text{GAE}}\, \tfrac{1}{N}\sum_i z_i, \tag{9}$$

where

$$B = \tfrac{1}{N}\textstyle\sum_i z_i z_i^\top \;-\; \big(\tfrac{1}{N}\sum_i z_i\big)\big(\tfrac{1}{N}\sum_i z_i\big)^\top \;+\; \lambda_{\text{pres}}\, I, \tag{10}$$

$$C = \tfrac{1}{N}\textstyle\sum_i h_i z_i^\top \;-\; \big(\tfrac{1}{N}\sum_i h_i\big)\big(\tfrac{1}{N}\sum_i z_i\big)^\top \;+\; \lambda_{\text{pres}}\, \widetilde{W}_{\text{dec}}(T^\star). \tag{11}$$

The Step 2 solution applies regularization strength λ_{pres} to decoder mass inside the OOD-active subspace and the larger strength $\lambda_{\text{pres}} + \lambda_{\text{geom}}$ outside it, so *any decoder mass that drifts off the Step 1 alignment is automatically shrunk back*. The full derivation is in Appendix A.5.

Applying GAE at inference. At inference, the ID-trained encoder remains unchanged: it extracts sparse feature activations z from each OOD example. Using the refitted decoder, GAE (Algorithm 1) reconstructs $\hat{h} = W_{\text{dec}}^{\text{GAE}}\, z + b_{\text{dec}}^{\text{GAE}}$. To control the effect of decoder refitting, we interpolate its output with the Step 1 dictionary:

$$W_{\text{dec}} = (1 - \alpha)\, W_{\text{dec}}^{\text{GAE}} + \alpha\, \widetilde{W}_{\text{dec}}(T^\star), \tag{12}$$

where $\alpha \in [0, 1]$ is the weight of the Step 1 dictionary, selected on held-out OOD activations (Appendix D.4). For $\alpha = 0$, GAE uses the refitted decoder. For $\alpha = 1$, Step 1-Only directly uses $W_{\text{dec}} = \widetilde{W}_{\text{dec}}(T^\star)$ and skips Step 2. This applies identically to both SAEs and transcoders.

4.3 Projection-Loss Bound for GAE

GAE's projection loss remains close to the optimum over rank-r subspaces when subspace estimation is accurate and decoder refitting introduces little additional misalignment. Step 1 aligns the dictionary with the estimated OOD-active subspace, while Step 2 can alter this alignment.

We quantify the subspace estimation error by $\varepsilon_N := \|\widehat{\Pi}_{\text{OOD}} - \Pi_{\text{OOD}}\|_F$. For decoder refitting, define $R_\perp := (I - \widehat{\Pi}_{\text{OOD}})C$ using C from Equation (11). Equation (8) shows that $R_\perp$ drives the decoder component outside this subspace. The Step 2 penalties limit this component, and interpolation scales it by $1-\alpha$. Under the rank condition below, this control bounds the additional subspace misalignment. Combining this bound with the subspace estimation error and Proposition 2 gives GAE's excess projection-loss bound.

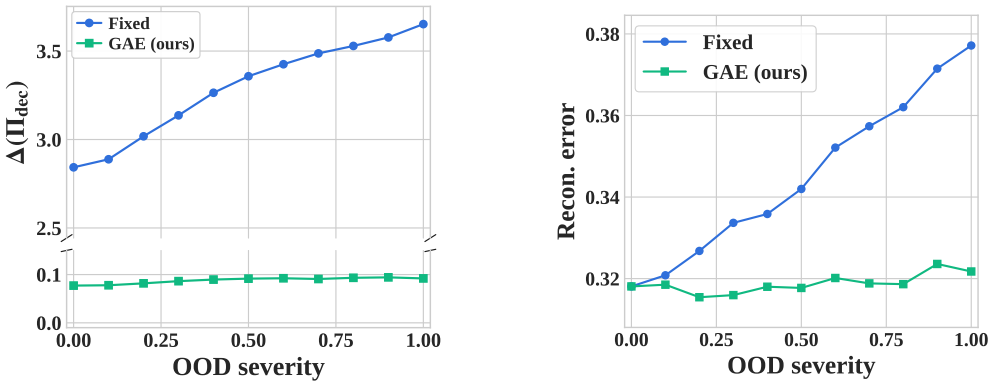
Corollary 1 (Projection Loss of GAE). *Let $\beta_{\text{OOD}} := \lambda_1(M_{\text{OOD}}) - \lambda_d(M_{\text{OOD}})$, and write $\sigma_r(\cdot)$ for the r-th singular value. Suppose $\gamma_{\text{OOD}} > 0$, $\operatorname{rank}(W_{\text{dec}}^{\text{ID}}) \geq r$, and $0 \leq \alpha < 1$. Let $\lambda_{\text{pres}}, \lambda_{\text{geom}} \geq 0$ with $\lambda_{\text{pres}} + \lambda_{\text{geom}} > 0$, and suppose B is invertible. For the dictionary W_{dec} in Equation (12) and its top-r projector Π_{dec}, assume $\sigma_r(\widehat{\Pi}_{\text{OOD}} W_{\text{dec}}) > 0$. Then*

$$\mathcal{L}_{\text{OOD}}(\Pi_{\text{dec}}) - \mathcal{L}_{\text{OOD}}(\Pi_{\text{OOD}}) \;\leq\; \frac{\beta_{\text{OOD}}}{2}\left(\varepsilon_N + \frac{2(1-\alpha)\|R_\perp\|_F}{(\lambda_{\text{pres}} + \lambda_{\text{geom}})\, \sigma_r(\widehat{\Pi}_{\text{OOD}} W_{\text{dec}})}\right)^2. \tag{13}$$

Thus, GAE's projection loss can stay close to the optimum even after refitting, provided both subspace estimation error and additional misalignment remain small. For $\alpha = 1$, Step 1-Only skips Step 2 and has projector $\widehat{\Pi}_{\text{OOD}}$, giving excess projection loss at most $\beta_{\text{OOD}}\varepsilon_N^2/2$. When $\varepsilon_N = 0$, this is the Step 1 endpoint of Theorem 1. The proof is given in Appendix A.7.

5 Experiments

We evaluate whether GAE restores explanation faithfulness under distribution shift. Section 5.1 tests the geometric mechanism in a controlled setting, and Section 5.2 evaluates on language models.



(a) Faithfulness gap.

(b) Reconstruction error.

Figure 2: **Controlled experiment on a toy MLP with OOD severity varied from 0 (ID) to 1 (maximum shift).** (a) The Fixed explainer's faithfulness gap $\Delta(\Pi_{\text{dec}})$ grows monotonically. (b) Its reconstruction error rises accordingly. GAE maintains near-zero gap and flat error throughout.

5.1 Controlled Experiment

We first test whether the geometric mechanism from Section 3 holds in a controlled setting. We train a 2-layer ReLU MLP with hidden dim $d=256$ and output dim $p=8$, and a linear-decoder SAE on its ID hidden activations, then continuously increase OOD severity by rotating and rescaling the input covariance (details in Appendix B.1). Figure 2 confirms that as severity increases, the faithfulness gap and reconstruction error of the Fixed explainer (the ID-trained explainer, used without adaptation) grow, while GAE closes the gap to near zero and keeps reconstruction error nearly flat.

5.2 Experiments on Language Models

5.2.1 Setup

Target models and explainers. We use two main frozen pretrained language models: GPT-2 Small [39] and Pythia-1.4B [40]. For each model, we train transcoders [5] and Top-K SAEs [7], both with dictionary size $k{=}32d$, as dictionary-based explainers (Equation (1)).

OOD settings. We consider three categories of distribution shift: **temporal** (FineWeb [41], web text collected after each model's pretraining cutoff), **domain** (Edgar [42], financial filings whose specialized vocabulary and structure differ from general web text), and **adversarial** (HaluEval [43], hallucination-inducing prompts that elicit atypical hidden representations). All three induce measurable second-moment shift in hidden activations, as verified in Appendix B.4.

Baselines. We compare training-free and training-based approaches. **Fixed** applies the ID-trained explainer without adaptation. **TERM** [16, 17] trains the ID explainer with tilted ERM to upweight tail samples. Among training-based methods, **Finetune** [44] warm-starts from the ID explainer on OOD activations, **Retrain** trains from scratch on OOD data, **SAEBoost** [18] adds a residual booster on OOD reconstruction residuals, and **FaithfulSAE** [15] retrains on the model's own generations. **GAE (ours)** is training-free: it uses only unlabeled OOD activations with no gradient computation. Detailed descriptions are in Appendix D.2.

Evaluation metrics. We evaluate causal faithfulness and reconstruction quality using three metrics. **Normalized AOPC (nAOPC)** [23] averages the normalized logit drop across multiple feature budgets when top-m features are removed (↑ is better). **Normalized comprehensiveness (nComp)** [24] measures the normalized logit drop at a single budget $m^*{=}32$ (↑ is better). Both measure the logit-level effect of ablating top features. **Delta cross-entropy (ΔCE)** [7] measures reconstruction quality: the cross-entropy change when activations are replaced with the explainer's reconstruction (≈ 0 is better). GAE optimizes a geometric objective (the faithfulness gap); improvements on the two causal metrics test whether geometric realignment yields faithfulness gains. Formal definitions are in Appendix D.5.

As shown in Table 1, all training-based baselines require gradient-based optimization over millions of tokens. Even the lightest, Finetune, processes 5M tokens over several minutes; Retrain, SAEBoost, and FaithfulSAE each consume 100M tokens and take tens of minutes to hours on a single GPU. *GAE requires no gradient computation at all*: one closed-form adaptation run completes in 0.5 s for GPT-2 and 2.9 s for Pythia-1.4B, using only $\sim$2,048 unlabeled OOD activations.

Table 1: **Computational cost per method.** We report token counts and wall-clock times. GAE takes under 3 s per model without training.

Method	Tokens (Vectors)	GPT-2	P-1.4B
Finetune	5M	$\sim$2 min	$\sim$12 min
Retrain	100M	$\sim$39 min	$\sim$4 hrs
SAEBoost	100M	$\sim$39 min	$\sim$4 hrs
FaithfulSAE	100M	$\sim$39 min	$\sim$4 hrs
GAE	**(2K)**	**0.5 s**	**2.9 s**

5.2.2 Faithfulness Results

Table 2: **Faithfulness under distribution shift (Transcoder, two models $\times$ three OOD settings).** GAE is training-free yet leads in all nine columns on GPT-2 and achieves the best nComp and $|\Delta\mathrm{CE}|$ in 5 of 6 Pythia-1.4B columns. **Bold**: best per column. Underline: second best.

	FineWeb (Temporal)			Edgar (Domain)			HaluEval (Adversarial)		
Method	**nAOPC**$\uparrow$	**nComp**$\uparrow$	$\lvert\Delta\mathbf{CE}\rvert\downarrow$	**nAOPC**$\uparrow$	**nComp**$\uparrow$	$\lvert\Delta\mathbf{CE}\rvert\downarrow$	**nAOPC**$\uparrow$	**nComp**$\uparrow$	$\lvert\Delta\mathbf{CE}\rvert\downarrow$
GPT-2 Small									
Fixed	0.857	1.017	0.0281	0.975	1.025	0.0201	0.735	0.737	0.0473
TERM	0.853	0.993	0.0283	0.964	0.999	0.0198	0.730	0.732	0.0523
Finetune	0.856	0.984	0.0172	0.971	1.117	0.0047	0.579	0.579	0.0105
Retrain	0.895	1.118	0.0218	0.936	1.034	0.0015	0.521	0.528	0.2763
SAEBoost	0.958	1.476	0.0177	0.979	1.542	0.0072	0.840	0.921	0.0212
FaithfulSAE	0.936	1.186	0.0213	0.976	1.134	0.0197	0.738	0.740	0.0586
GAE	**0.960**	**1.494**	**0.0167**	**0.981**	**1.618**	**0.0009**	**0.871**	**0.963**	**0.0014**
Pythia-1.4B									
Fixed	0.839	1.091	0.0278	0.725	0.828	0.0300	0.899	1.021	0.0354
TERM	0.845	1.043	0.0271	0.895	0.981	0.0298	0.896	1.104	0.0349
Finetune	0.859	1.249	**0.0264**	0.684	0.860	0.0282	0.925	1.502	0.0280
Retrain	0.894	1.315	0.0405	0.746	0.821	0.0329	0.894	1.393	0.0305
SAEBoost	0.908	1.296	0.0284	0.903	1.297	0.0296	0.965	1.530	0.0283
FaithfulSAE	0.858	1.056	0.0272	0.724	0.822	0.0323	0.899	1.171	0.0307
GAE	**0.915**	**1.354**	0.0269	**0.988**	**1.652**	**0.0230**	**0.968**	**1.693**	**0.0276**

Table 2 reports faithfulness across three OOD settings for GPT-2 Small and Pythia-1.4B (Transcoder). Despite using no gradient updates, *GAE leads on all three metrics for GPT-2 across every OOD setting*, surpassing training-based baselines that consume up to 100M tokens (cf. Table 1). On adversarial shift, GAE improves nComp over the strongest training-based baseline on this metric, SAEBoost, by 4.6% (0.963 vs 0.921). Relative to the same baseline, GAE reduces $|\Delta\mathrm{CE}|$ by 93% (0.0014 vs 0.0212). On Pythia-1.4B, GAE achieves the best nAOPC and nComp on all three settings; $|\Delta\mathrm{CE}|$ is best on Edgar and HaluEval but slightly elevated on FineWeb (0.027 vs Finetune's 0.026), where the more diffuse eigenvalue decay weakens the rank-r approximation.

Table 3 reports the same evaluation for Top-K SAEs. On GPT-2, GAE again surpasses every training-based baseline on all nine columns, including gains on Edgar (nComp 0.809 vs Retrain's 0.797) and HaluEval ($|\Delta\mathrm{CE}|$ 0.0017 vs Finetune's 0.0218). On Pythia-1.4B, Retrain is a stronger competitor than for transcoders, taking best nComp on FineWeb (1.681 vs 1.677) and best nAOPC on Edgar (0.970 vs 0.968). GAE still leads on nComp and $|\Delta\mathrm{CE}|$ in 5 of 6 columns and achieves the best $|\Delta\mathrm{CE}|$ across all three settings. The pattern is consistent: *GAE matches or surpasses methods that require orders of magnitude more computation.*

Comparison with geometric adaptation. We compare geometric adaptation methods to show that GAE's decoder adaptation for causal faithfulness outperforms generic geometric alignment under OOD shift. SA [30] aligns source and target PCA subspaces, whereas CORAL [31] matches covariance statistics. Both decoder adaptations keep the sparse encoder frozen. They and dense PCA use same unlabeled OOD activations and evaluation protocol. GAE leads on nAOPC, nComp, and $|\Delta\mathrm{CE}|$ (Table 4). Appendix D.3 also reports the result on SAE.

Table 3: **Faithfulness under distribution shift (SAE, two models × three OOD settings).** GAE leads in all nine columns on GPT-2 and leads on nComp and $|\Delta\text{CE}|$ in 5 of 6 Pythia-1.4B columns. **Bold**: best. Underline: second best.

	FineWeb (Temporal)			Edgar (Domain)			HaluEval (Adversarial)		
Method	**nAOPC↑**	**nComp↑**	**$\|\Delta\text{CE}\|$↓**	**nAOPC↑**	**nComp↑**	**$\|\Delta\text{CE}\|$↓**	**nAOPC↑**	**nComp↑**	**$\|\Delta\text{CE}\|$↓**
GPT-2 Small									
Fixed	0.735	0.796	0.0185	0.650	0.667	0.0089	0.930	1.134	0.0406
TERM	0.741	0.790	0.0183	0.655	0.682	0.0145	0.898	0.987	0.0401
Finetune	0.725	0.802	0.0015	0.658	0.687	0.0068	0.932	1.155	0.0218
Retrain	0.766	0.856	0.0375	0.715	0.797	0.0065	0.930	1.276	0.0300
SAEBoost	0.704	0.786	0.0120	0.604	0.650	0.0098	0.909	1.243	0.0448
FaithfulSAE	0.725	0.760	0.0262	0.657	0.683	0.0074	0.908	1.082	0.0278
GAE	**0.768**	**0.871**	**0.0011**	**0.723**	**0.809**	**0.0037**	**0.953**	**1.303**	**0.0017**
Pythia-1.4B									
Fixed	0.962	1.536	0.0216	0.953	1.690	0.0170	0.985	1.425	0.0252
TERM	0.965	1.486	0.0237	0.959	1.787	0.0167	0.984	1.762	0.0292
Finetune	0.963	1.618	0.0216	0.959	1.832	0.0131	0.988	1.880	0.0174
Retrain	0.983	**1.681**	0.0563	**0.970**	1.848	0.0261	**1.000**	1.855	0.0209
SAEBoost	0.971	1.451	0.0211	0.955	1.598	0.0152	**1.000**	1.830	0.0166
FaithfulSAE	0.982	1.642	0.0307	0.953	1.769	0.0205	**1.000**	1.678	0.0514
GAE	**0.985**	1.677	**0.0207**	0.968	**1.946**	**0.0098**	**1.000**	**1.885**	**0.0163**

Table 4: **Geometric adaptation on HaluEval (Transcoder).** GAE improves causal faithfulness and reconstruction over PCA, SA, and CORAL without gradient updates. **Bold**: best. Underline: second best.

Method	**GPT-2 Small**			**Pythia-1.4B**		
	nAOPC↑	**nComp↑**	**$\|\Delta\text{CE}\|$↓**	**nAOPC↑**	**nComp↑**	**$\|\Delta\text{CE}\|$↓**
Fixed	0.735	0.737	0.0473	0.899	1.021	0.0354
PCA	0.784	0.910	0.0808	0.942	1.364	0.0581
SA	0.801	0.895	0.0415	0.536	0.544	0.0816
CORAL	0.405	0.409	0.0580	0.931	1.305	0.1048
GAE	**0.871**	**0.963**	**0.0014**	**0.968**	**1.693**	**0.0276**

Table 5: **Llama-3.1-8B SAE on HaluEval.** GAE leads on all metrics without gradient updates. **Bold**: best. Underline: second best.

Method	**nAOPC↑**	**nComp↑**	**$\|\Delta\text{CE}\|$↓**
Fixed	0.5719	0.7892	0.0652
Finetune	0.5906	0.7913	0.0609
SAEBoost	0.5828	0.7902	0.0631
GAE	**0.5972**	**0.7936**	**0.0602**

Larger-model evaluation. We evaluate Llama-3.1-8B with a public LlamaScope 32× Top-K SAE [45] containing 128K features on HaluEval. Using $r = 64$ and $N = 2{,}048$ unlabeled OOD activation vectors, GAE adapts the decoder while keeping the model and dictionary encoder frozen. Table 5 shows higher nAOPC and nComp and lower $|\Delta\text{CE}|$ than Fixed, Finetune, and SAEBoost. Appendix H provides the setup and computational cost details.

5.2.3 Case Study: Circuit Attribution under Distribution Shift

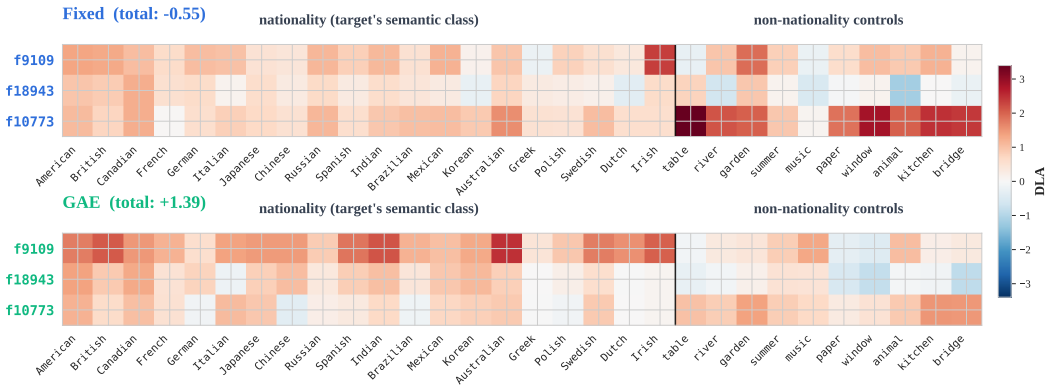
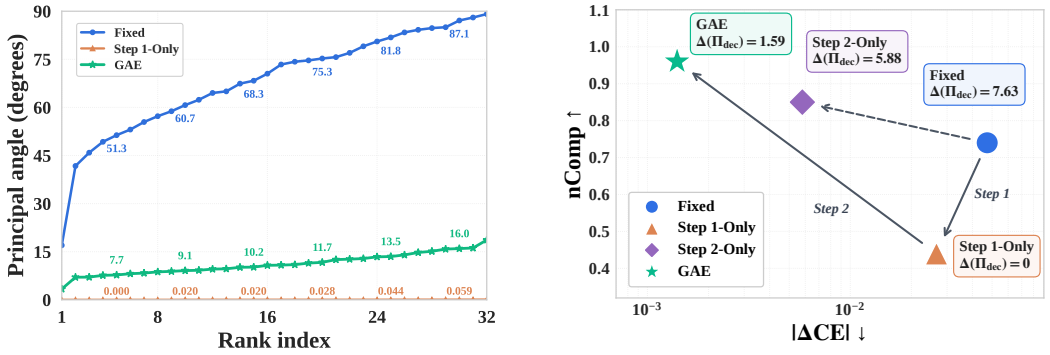


Figure 3: **Per-feature DLA on a prompt predicting 'American' (GPT-2, Transcoder).** Both methods share the same encoder and top-3 features; only the decoder columns differ. Each cell shows a feature's direct logit attribution (DLA) to nationality tokens (left, 20 tokens) vs. non-nationality controls (right, 10 tokens). Fixed's total class-specificity is -0.55 (circuit points away from the target class); GAE's is $+1.39$ (circuit points toward it).

For a prompt predicting ‘ American’, we compare the same feature IDs under Fixed and GAE, with identical activations. Direct logit attribution (DLA) measures a feature’s contribution to a token’s logit. For each feature, we subtract its mean DLA on 10 non-nationality controls from its mean DLA on 20 nationality tokens. Class-specificity sums these differences over the three selected features.

Fixed scores -0.55, whereas GAE scores $+1.39$ (Figure 3). For these selected features, decoder adaptation increases attribution to the target class while their activations remain unchanged. Appendix E gives the protocol and an additional first-name case study.

5.2.4 Mechanism Analysis



(a) Subspace alignment.

(b) Step ablation.

Figure 4: **Mechanism analysis (GPT-2, Transcoder). (a)** Sorted principal angles between each explainer’s top-r subspace and $\widehat{\Pi}_{\mathrm{OOD}}$. Step 1-Only yields near-zero angles, while GAE retains residual angles after decoder refitting. **(b)** Step ablation on HaluEval: Step 1-Only closes the faithfulness gap to 0 yet drops nComp from 0.74 to 0.44. Step 2-Only (dashed arrow) leaves a gap of 5.88, whereas GAE achieves a gap of 1.59, higher nComp, and lower $|\Delta\mathrm{CE}|$. This comparison supports aligning the dictionary preservation reference before refitting the decoder.

Subspace alignment. Figure 4(a) measures the principal angles between each explainer’s top-r decoder subspace and the OOD-active subspace $\widehat{\Pi}_{\mathrm{OOD}}$. Step 1-Only yields near-zero angles, while GAE retains residual angles after decoder refitting. Both are better aligned with $\widehat{\Pi}_{\mathrm{OOD}}$ than Fixed. Appendix C illustrates approximately linear Step 1 projection-loss improvement with $\Delta(\Pi_{\mathrm{ID}})^2$.

Step ablation. Figure 4(b) ablates each step of GAE on HaluEval. Step 2-Only skips Step 1 and uses the Fixed ID dictionary as the dictionary preservation reference. The encoder, data, objective, hyperparameters, and evaluation protocol are unchanged. Step 1-Only closes the faithfulness gap from 7.63 to 0 by construction, yet nComp drops from 0.74 to 0.44, showing that subspace alignment alone does not ensure higher nComp. Step 2-Only improves nComp to 0.85 and reduces $|\Delta\mathrm{CE}|$ to 0.0058, but leaves a faithfulness gap of 5.88. In contrast, GAE achieves a smaller gap (1.59), the highest nComp (0.96), and the lowest $|\Delta\mathrm{CE}|$ (0.0014). These results support complementary roles: Step 1 provides an OOD-aligned dictionary as the preservation reference, and Step 2 improves its compatibility with the frozen sparse feature activations. Corollary 1 gives the corresponding projection-loss bound for GAE. See Appendix F for sensitivity.

6 Conclusion

We showed that OOD faithfulness degradation in dictionary-based explainers has a geometric cause: the decoder subspace drifts from the directions the model actively uses. The faithfulness gap $\Delta(\Pi_{\mathrm{dec}})$ formalizes this misalignment and provably controls the reducible part of OOD projection loss. GAE targets the gap with a closed-form subspace rotation and decoder refit, using only unlabeled OOD activations while preserving the original sparse feature activations. Across three shift types on GPT-2 Small and Pythia-1.4B, GAE outperforms all training-based baselines on 5 of 6 settings, completing in under 3 seconds without any gradient computation. An additional evaluation on Llama-3.1-8B with a public 128K-feature SAE improves all three reported metrics on HaluEval, with adaptation taking approximately 30 seconds. GAE also relies on a top-r SVD truncation of $W_{\mathrm{dec}}^{\mathrm{ID}}$, so any feature information carried in the residual $(d-r)$ singular directions is dropped before adaptation. Extending GAE to adaptive rank selection, encoder adaptation, and connections to optimal transport on the Grassmannian [46] are promising future directions.

Acknowledgements

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00457216). This research was supported by a fund from the research program of the Korea Medical Institute.(KMI20250213) This work was supported by National Institute for Mathematical Sciences (NIMS) grant funded by the Korea government (MSIT) (No. S26120000).

References

[1] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, et al. Causal abstraction: A theoretical foundation for mechanistic interpretability. Journal of Machine Learning Research, 26(83):1–64, 2025.

[2] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. arXiv preprint arXiv:2209.11895, 2022.

[3] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. arXiv preprint arXiv:2501.16496, 2025.

[4] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

[5] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable llm feature circuits. Advances in Neural Information Processing Systems, 37:24375–24410, 2024.

[6] Adly Templeton. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Anthropic, 2024.

[7] Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. arXiv preprint arXiv:2406.04093, 2024.

[8] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. arXiv preprint arXiv:2403.19647, 2024.

[9] Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? arXiv preprint arXiv:2004.03685, 2020.

[10] Google DeepMind Safety Research. Negative results for sparse autoencoders on downstream tasks and deprioritising sae research. DeepMind Safety Research Blog, 2025. Blog post.

[11] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. Advances in neural information processing systems, 31, 2018.

[12] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 3681–3688, 2019.

[13] Chris Lin, Ian Covert, and Su-In Lee. On the robustness of removal-based feature attributions. Advances in Neural Information Processing Systems, 36:79613–79666, 2023.

[14] Chiara Balestra, Bin Li, and Emmanuel Müller. On the consistency and robustness of saliency explanations for time series classification. arXiv preprint arXiv:2309.01457, 2023.

[15] Seonglae Cho, Harryn Oh, Donghyun Lee, Luis Rodrigues Vieira, Andrew Bermingham, and Ziad El Sayed. Faithfulsae: Towards capturing faithful features with sparse autoencoders without external datasets dependency. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 297–314, 2025.

[16] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. arXiv preprint arXiv:2007.01162, 2020.

[17] Aashiq Muhamed, Mona Diab, and Virginia Smith. Decoding dark matter: Specialized sparse autoencoders for interpreting rare concepts in foundation models. In Findings of the Association for Computational Linguistics: NAACL 2025, pages 1604–1635, 2025.

[18] Nikita Koriagin, Yaroslav Aksenov, Daniil Laptev, Gleb Gerasimov, Nikita Balagansky, and Daniil Gavrilov. Teach old saes new domain tricks with boosting. arXiv preprint arXiv:2507.12990, 2025.

[19] Leonard Bereska and Efstratios Gavves. Mechanistic interpretability for ai safety–a review. arXiv preprint arXiv:2404.14082, 2024.

[20] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems, 31, 2018.

[21] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. arXiv preprint arXiv:2309.08600, 2023.

[22] Gonçalo Paulo, Stepan Shabalin, and Nora Belrose. Transcoders beat sparse autoencoders for interpretability. arXiv preprint arXiv:2501.18823, 2025.

[23] Joakim Edin, Andreas Geert Motzfeldt, Casper L Christensen, Tuukka Ruotsalo, Lars Maaløe, and Maria Maistro. Normalized aopc: Fixing misleading faithfulness metrics for feature attributions explainability. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1715–1730, 2025.

[24] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. Eraser: A benchmark to evaluate rationalized nlp models. In Proceedings of the 58th annual meeting of the association for computational linguistics, pages 4443–4458, 2020.

[25] Lawrence Chan, Adria Garriga-Alonso, Nicholas Goldowsky-Dill, Ryan Greenblatt, Jenny Nitishinskaya, Ansh Radhakrishnan, Buck Shlegeris, and Nate Thomas. Causal scrubbing: A method for rigorously testing interpretability hypotheses. In AI Alignment Forum, volume 2, 2022.

[26] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. Advances in neural information processing systems, 35:17359–17372, 2022.

[27] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In International conference on machine learning, pages 3519–3529. PMLR, 2019.

[28] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. Advances in neural information processing systems, 30, 2017.

[29] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. Advances in neural information processing systems, 34:4738–4750, 2021.

[30] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In Proceedings of the IEEE international conference on computer vision, pages 2960–2967, 2013.

[31] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In Proceedings of the AAAI conference on artificial intelligence, 2016.

[32] Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nature Communications, 12(1):4122, 2021.

[33] Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. Intrinsic dimension of data representations in deep neural networks. Advances in Neural Information Processing Systems, 32, 2019.

[34] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers), pages 7319–7328, 2021.

[35] Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. In Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 278–300, 2024.

[36] Callum McDougall, Arthur Conmy, János Kramár, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. Gemma scope 2: Technical paper. Technical report, Google DeepMind, 2025.

[37] Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. arXiv preprint arXiv:2407.14435, 2024.

[38] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. Psychometrika, 31(1):1–10, 1966.

[39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9, 2019.

[40] Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In International Conference on Machine Learning, pages 2397–2430. PMLR, 2023.

[41] Guilherme Penedo, Hynek Kydlícek, Loubna Ben Allal, and Thomas Wolf. Fineweb: decanting the web for the finest text data at scale. HuggingFace. Accessed: Jul, 12, 2024.

[42] Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. Edgar-corpus: Billions of tokens make the world go round. In Proceedings of the Third Workshop on Economics and Natural Language Processing, pages 13–18, 2021.

[43] Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In Proceedings of the 2023 conference on empirical methods in natural language processing, pages 6449–6464, 2023.

[44] Connor Kissane, Robert Krzyzanowski, Arthur Conmy, and Neel Nanda. Saes (usually) transfer between base and chat models. Alignment Forum, 2024. URL https://www.alignmentforum.org/posts/fmwk6qxrpW8d4jvbd/saes-usually-transfer-between-base-and-chat-models.

[45] Zhengfu He, Wentao Shu, Xuyang Ge, Lingjie Chen, Junxuan Wang, Yunhua Zhou, Frances Liu, Qipeng Guo, Xuanjing Huang, Zuxuan Wu, et al. Llama scope: Extracting millions of features from llama-3.1-8b with sparse autoencoders. arXiv preprint arXiv:2410.20526, 2024.

[46] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. SIAM journal on Matrix Analysis and Applications, 20(2):303–353, 1998.

[47] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. http://Skylion007.github.io/OpenWebTextCorpus, 2019.

[48] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. arXiv preprint arXiv:2101.00027, 2020.

[49] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. arXiv preprint arXiv:2202.10054, 2022.

A Proofs and Derivations

A.1 Proof of Proposition 1

Setup. Write the second-moment shift as $E = M_{\mathrm{OOD}} - M_{\mathrm{ID}}$, so that $M_{\mathrm{OOD}} = M_{\mathrm{ID}} + E$. The projectors Π_{ID} and Π_{OOD} correspond to the top-r eigenspaces of M_{ID} and $M_{\mathrm{ID}} + E$, respectively. We wish to bound $\Delta(\Pi_{\mathrm{ID}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{ID}}\|_F$ in terms of $\|E\|_F$.

Step 1: Projector distance via principal angles. Let $\theta_1, \ldots, \theta_r$ be the principal angles between the column spaces of Π_{ID} and Π_{OOD}. For two rank-r orthogonal projectors, the Frobenius norm of their difference satisfies

$$\|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{ID}}\|_F^2 = 2\sum_{i=1}^{r} \sin^2\theta_i.$$

Step 2: Using the ID eigengap. The eigenbasis argument in Proposition 2, applied to M_{ID}, gives

$$\frac{\gamma_{\mathrm{ID}}}{2}\|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F^2 \le \mathrm{tr}((\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}})M_{\mathrm{ID}}),$$

where $\gamma_{\mathrm{ID}} = \lambda_r(M_{\mathrm{ID}}) - \lambda_{r+1}(M_{\mathrm{ID}})$ is the eigengap of M_{ID} at rank r.

Step 3: Using OOD optimality. Since Π_{OOD} maximizes the retained OOD energy, $\mathrm{tr}((\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}})M_{\mathrm{OOD}}) \le 0$. Using $M_{\mathrm{ID}} = M_{\mathrm{OOD}} - E$ and Frobenius Cauchy–Schwarz,

$$\begin{aligned}\mathrm{tr}((\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}})M_{\mathrm{ID}}) &\le -\mathrm{tr}((\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}})E)\\ &\le \|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F\|E\|_F.\end{aligned}$$

Step 4: Combining. Combining Steps 2 and 3 for nonzero projector distance,

$$\Delta(\Pi_{\mathrm{ID}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{ID}}\|_F = \sqrt{2\sum_{i=1}^{r}\sin^2\theta_i} \le \frac{2}{\gamma_{\mathrm{ID}}}\|E\|_F = \frac{2}{\gamma_{\mathrm{ID}}}\|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F.$$

If the projector distance is zero, the bound is immediate. □

A.2 Decomposition Equation (3)

Step 1: Expressing $\mathcal{L}_{\mathrm{OOD}}$ in terms of M_{OOD}. For any rank-r orthogonal $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$, the reconstruction error under P_{OOD} is

$$\begin{aligned}\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) &= \mathbb{E}_{X\sim P_{\mathrm{OOD}}}\|h(X) - \Pi_{\mathrm{dec}}\,h(X)\|_2^2\\ &= \mathbb{E}\|(I - \Pi_{\mathrm{dec}})\,h(X)\|_2^2\\ &= \mathbb{E}\ \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\,h(X)\,h(X)^\top\,(I - \Pi_{\mathrm{dec}})^\top\big]\\ &= \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\,M_{\mathrm{OOD}}\,(I - \Pi_{\mathrm{dec}})\big],\end{aligned} \tag{14}$$

where the third step uses $\|a\|_2^2 = \mathrm{tr}(aa^\top)$ and the fourth step swaps expectation and trace, with $M_{\mathrm{OOD}} = \mathbb{E}[h(X)h(X)^\top]$. This derivation explains why we use the uncentered second moment rather than the centered covariance: the projection objective depends directly on M_{OOD}, which captures covariance shift as well as mean shift. Since $I - \Pi_{\mathrm{dec}}$ is an orthogonal projector, it is idempotent: $(I - \Pi_{\mathrm{dec}})^2 = I - \Pi_{\mathrm{dec}}$. Using the cyclic property of trace,

$$\mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\,M_{\mathrm{OOD}}\,(I - \Pi_{\mathrm{dec}})\big] = \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})^2\,M_{\mathrm{OOD}}\big] = \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\,M_{\mathrm{OOD}}\big]. \tag{15}$$

Step 2: Deriving the decomposition. Applying the same identity with Π_{OOD} gives $\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) = \mathrm{tr}\big[(I - \Pi_{\mathrm{OOD}})\,M_{\mathrm{OOD}}\big]$. Taking the difference,

$$\begin{aligned}\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) &= \mathrm{tr}\big[(I - \Pi_{\mathrm{dec}})\,M_{\mathrm{OOD}}\big] - \mathrm{tr}\big[(I - \Pi_{\mathrm{OOD}})\,M_{\mathrm{OOD}}\big]\\ &= \mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}})\,M_{\mathrm{OOD}}\big].\end{aligned} \tag{16}$$

Rearranging gives the claimed decomposition:

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) + \mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}})\,M_{\mathrm{OOD}}\big].$$

Step 3: Nonnegativity and optimality. It remains to show that the second term is nonnegative. Let $M_{\mathrm{OOD}} = \sum_{i=1}^{d} \lambda_i \, u_i u_i^\top$ be the eigendecomposition with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Since Π_{OOD} projects onto the top-r eigenspace,

$$\mathrm{tr}(\Pi_{\mathrm{OOD}} \, M_{\mathrm{OOD}}) = \sum_{i=1}^{r} \lambda_i.$$

By Ky Fan's maximum principle, $\sum_{i=1}^{r} \lambda_i = \max_{\Pi_{\mathrm{dec}} \in \mathcal{C}_r} \mathrm{tr}(\Pi_{\mathrm{dec}} \, M_{\mathrm{OOD}})$. Therefore, for any $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$,

$$\mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}) \, M_{\mathrm{OOD}}\big] = \mathrm{tr}(\Pi_{\mathrm{OOD}} \, M_{\mathrm{OOD}}) - \mathrm{tr}(\Pi_{\mathrm{dec}} \, M_{\mathrm{OOD}}) \geq 0,$$

with equality if and only if Π_{dec} also projects onto a top-r eigenspace of M_{OOD}. This proves both the nonnegativity and the optimality $\Pi_{\mathrm{OOD}} \in \arg\min_{\Pi_{\mathrm{dec}} \in \mathcal{C}_r} \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})$. □

A.3 Proof of Proposition 2

From the decomposition (3), the explainer-dependent component equals $\mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}) \, M_{\mathrm{OOD}}\big]$. We derive both bounds by expanding this trace in the eigenbasis of M_{OOD}.

Setup. Let $M_{\mathrm{OOD}} = \sum_{i=1}^{d} \lambda_i \, u_i u_i^\top$ be the eigendecomposition with $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. The OOD-active subspace is $\Pi_{\mathrm{OOD}} = \sum_{i=1}^{r} u_i u_i^\top$, so $u_i^\top \Pi_{\mathrm{OOD}} \, u_i = \mathbf{1}_{i \leq r}$. For the explainer subspace $\Pi_{\mathrm{dec}} \in \mathcal{C}_r$, define $p_i := u_i^\top \Pi_{\mathrm{dec}} \, u_i \in [0, 1]$, the fraction of the i-th OOD eigendirection captured by Π_{dec}. Expanding the trace in this eigenbasis gives

$$\begin{aligned} \mathrm{tr}\big[(\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}) \, M_{\mathrm{OOD}}\big] &= \sum_{i=1}^{d} \lambda_i \, (u_i^\top \Pi_{\mathrm{OOD}} \, u_i - p_i) \\ &= \sum_{i=1}^{r} \lambda_i \, (1 - p_i) - \sum_{i=r+1}^{d} \lambda_i \, p_i. \end{aligned} \tag{17}$$

The first sum captures the OOD energy in the top-r directions that the explainer misses (since $1 - p_i$ is the fraction lost). The second sum captures the OOD energy in the bottom directions that the explainer unnecessarily covers.

Connecting to the faithfulness gap. For two rank-r subspaces, the Frobenius norm of their difference satisfies $\|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F^2 = 2 \sum_{i=1}^{r} (1 - p_i)$. Moreover, since both have the same rank, $\sum_{i=1}^{r} (1 - p_i) = \sum_{i=r+1}^{d} p_i$. Denoting $S := \sum_{i=1}^{r} (1 - p_i)$, we have

$$\Delta(\Pi_{\mathrm{dec}})^2 = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F^2 = 2S, \quad \text{and} \quad \sum_{i=r+1}^{d} p_i = S. \tag{18}$$

Upper bound. Using $\lambda_i \leq \lambda_1$ for $i \leq r$ and $\lambda_i \geq \lambda_d$ for $i > r$ in Equation (17),

$$\begin{aligned} \sum_{i=1}^{r} \lambda_i (1 - p_i) - \sum_{i=r+1}^{d} \lambda_i \, p_i &\leq \lambda_1 \sum_{i=1}^{r} (1 - p_i) - \lambda_d \sum_{i=r+1}^{d} p_i \\ &= \lambda_1 \, S - \lambda_d \, S = (\lambda_1 - \lambda_d) \, S = \frac{\lambda_1(M_{\mathrm{OOD}}) - \lambda_d(M_{\mathrm{OOD}})}{2} \, \Delta(\Pi_{\mathrm{dec}})^2. \end{aligned} \tag{19}$$

Lower bound. Using $\lambda_i \geq \lambda_r$ for $i \leq r$ and $\lambda_i \leq \lambda_{r+1}$ for $i > r$,

$$\begin{aligned} \sum_{i=1}^{r} \lambda_i (1 - p_i) - \sum_{i=r+1}^{d} \lambda_i \, p_i &\geq \lambda_r \sum_{i=1}^{r} (1 - p_i) - \lambda_{r+1} \sum_{i=r+1}^{d} p_i \\ &= \lambda_r \, S - \lambda_{r+1} \, S = (\lambda_r - \lambda_{r+1}) \, S = \frac{\gamma_{\mathrm{OOD}}}{2} \, \Delta(\Pi_{\mathrm{dec}})^2. \end{aligned} \tag{20}$$

□

For $\gamma_{\mathrm{ID}} > 0$, substituting Proposition 1 into the upper bound of Proposition 2 gives

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \leq \frac{\lambda_1(M_{\mathrm{OOD}}) - \lambda_d(M_{\mathrm{OOD}})}{2} \left(\frac{2}{\gamma_{\mathrm{ID}}} \|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F \right)^2 .$$

This bound is quadratic in the second-moment shift and does not require $\gamma_{\mathrm{OOD}} > 0$.

A.4 Derivation of the GAE Procrustes Solution (Step 1)

We derive the closed-form solution to Equation (6). Substituting $\widetilde{W}_{\mathrm{dec}}(T) = U_{\mathrm{OOD}}^{(:r)} T U_{\mathrm{dec}}^{\top} W_{\mathrm{dec}}^{\mathrm{ID}}$ and expanding:

$$\|\widetilde{W}_{\mathrm{dec}}(T) - W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2 = \|W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2 + \|U_{\mathrm{dec}}^{\top} W_{\mathrm{dec}}^{\mathrm{ID}}\|_F^2 - 2\,\mathrm{tr}\big[(W_{\mathrm{dec}}^{\mathrm{ID}})^{\top} U_{\mathrm{OOD}}^{(:r)} T U_{\mathrm{dec}}^{\top} W_{\mathrm{dec}}^{\mathrm{ID}}\big], \quad (21)$$

where we used $T^{\top}T = I$ and $(U_{\mathrm{OOD}}^{(:r)})^{\top} U_{\mathrm{OOD}}^{(:r)} = I$. The first two terms are independent of T, so minimization reduces to

$$T^{\star} = \arg\max_{T \in \mathcal{O}_r} \mathrm{tr}(G\,T), \qquad G = U_{\mathrm{dec}}^{\top} W_{\mathrm{dec}}^{\mathrm{ID}} (W_{\mathrm{dec}}^{\mathrm{ID}})^{\top} U_{\mathrm{OOD}}^{(:r)}.$$

Let $G = \widetilde{U}\Sigma\widetilde{V}^{\top}$ be the SVD of G. Setting $R = \widetilde{V}^{\top} T \widetilde{U}$, we have $\mathrm{tr}(G\,T) = \mathrm{tr}(\Sigma\,R) = \sum_i \sigma_i R_{ii}$. Since $R \in \mathcal{O}_r$, each $|R_{ii}| \leq 1$, so the maximum is achieved at $R = I_r$, giving $T^{\star} = \widetilde{V}\widetilde{U}^{\top}$. □

The same orthonormality identities give the Step 1 Gram identity

$$\widetilde{W}_{\mathrm{dec}}(T)^{\top} \widetilde{W}_{\mathrm{dec}}(T) = (W_{\mathrm{dec}}^{\mathrm{ID}})^{\top} U_{\mathrm{dec}} U_{\mathrm{dec}}^{\top} W_{\mathrm{dec}}^{\mathrm{ID}}. \quad (22)$$

Thus Step 1 preserves the pairwise inner products of the ID dictionary columns projected onto $\mathrm{span}(U_{\mathrm{dec}})$. Step 2 and interpolation can change this rank-r Gram geometry.

A.5 Derivation of the Step 2 Closed-Form Solution

Given paired target activations h_i and fixed sparse feature activations z_i, Step 2 solves the following convex quadratic objective over $W_{\mathrm{dec}} \in \mathbb{R}^{d \times k}$ and $b \in \mathbb{R}^d$:

$$\min_{W_{\mathrm{dec}},\, b} \; \frac{1}{N} \sum_{i=1}^{N} \|h_i - W_{\mathrm{dec}}\, z_i - b\|^2 \;+\; \lambda_{\mathrm{geom}} \,\|(I - \widehat{\Pi}_{\mathrm{OOD}})\, W_{\mathrm{dec}}\|_F^2 \;+\; \lambda_{\mathrm{pres}} \,\|W_{\mathrm{dec}} - \widetilde{W}_{\mathrm{dec}}(T^{\star})\|_F^2.$$

Bias. Setting $\partial/\partial b = 0$ gives $b^{\star} = \frac{1}{N}\sum_i h_i - W_{\mathrm{dec}}^{\star} \frac{1}{N} \sum_i z_i$, i.e., Equation (9).

Decoder. Substituting $b^{\star}$ centers the data: define $h_i^c = h_i - \frac{1}{N}\sum_j h_j$ and $z_i^c = z_i - \frac{1}{N}\sum_j z_j$. The problem reduces to

$$\min_{W_{\mathrm{dec}}} \; \frac{1}{N} \sum_{i=1}^{N} \|h_i^c - W_{\mathrm{dec}}\, z_i^c\|^2 \;+\; \lambda_{\mathrm{geom}} \,\|(I - \widehat{\Pi}_{\mathrm{OOD}})\, W_{\mathrm{dec}}\|_F^2 \;+\; \lambda_{\mathrm{pres}} \,\|W_{\mathrm{dec}} - \widetilde{W}_{\mathrm{dec}}(T^{\star})\|_F^2.$$

Setting $\partial/\partial W_{\mathrm{dec}} = 0$:

$$\big[\lambda_{\mathrm{pres}}\, I + \lambda_{\mathrm{geom}}(I - \widehat{\Pi}_{\mathrm{OOD}})\big]\, W_{\mathrm{dec}} \;+\; W_{\mathrm{dec}}\, \tfrac{1}{N}\textstyle\sum_i z_i^c z_i^{c\top} \;=\; \tfrac{1}{N}\textstyle\sum_i h_i^c z_i^{c\top} \;+\; \lambda_{\mathrm{pres}}\, \widetilde{W}_{\mathrm{dec}}(T^{\star}).$$

The left-hand coefficient $\Lambda := \lambda_{\mathrm{pres}}\, I + \lambda_{\mathrm{geom}}(I - \widehat{\Pi}_{\mathrm{OOD}})$ is diagonal in the OOD basis, with eigenvalue λ_{pres} on $\mathrm{span}(U_{\mathrm{OOD}}^{(:r)})$ and $\lambda_{\mathrm{pres}} + \lambda_{\mathrm{geom}}$ on its complement. Denoting $B = \frac{1}{N}\sum_i z_i^c z_i^{c\top} + \lambda_{\mathrm{pres}}\, I$ and $C = \frac{1}{N}\sum_i h_i^c z_i^{c\top} + \lambda_{\mathrm{pres}}\, \widetilde{W}_{\mathrm{dec}}(T^{\star})$, the system decouples row-wise in the OOD basis into two standard ridge regressions:

- **Inside** $\widehat{\Pi}_{\mathrm{OOD}}$ (first r rows): $W_{\mathrm{in}} = C_{\mathrm{in}}\, B^{-1}$, with ridge level λ_{pres}.
- **Outside** $\widehat{\Pi}_{\mathrm{OOD}}$ (remaining $d-r$ rows): $W_{\mathrm{out}} = C_{\mathrm{out}}\, (B + \lambda_{\mathrm{geom}}\, I)^{-1}$, with ridge level $\lambda_{\mathrm{pres}} + \lambda_{\mathrm{geom}}$.

Combining via the projector yields Equation (8): $W_{\mathrm{dec}}^{\mathrm{GAE}} = \widehat{\Pi}_{\mathrm{OOD}}\, C\, B^{-1} + (I - \widehat{\Pi}_{\mathrm{OOD}})\, C\, (B + \lambda_{\mathrm{geom}}\, I)^{-1}$. The centered feature covariance has rank at most $N - 1$. For $N \leq k$, $\lambda_{\mathrm{pres}} > 0$ is therefore required for B to be invertible and ensures that both inverse matrices exist. □

A.6 Proof of Theorem 1

Proof. By Equation (5), the rank-r dictionary $\widetilde{W}_{\mathrm{dec}}(T^\star)$ has column space $\mathrm{span}(U_{\mathrm{OOD}}^{(:r)})$, with projector

$$\widehat{\Pi}_{\mathrm{OOD}} = U_{\mathrm{OOD}}^{(:r)}(U_{\mathrm{OOD}}^{(:r)})^\top. \tag{23}$$

Under the condition $\widehat{\Pi}_{\mathrm{OOD}} = \Pi_{\mathrm{OOD}}$, this gives $\Delta(\widehat{\Pi}_{\mathrm{OOD}}) = \|\Pi_{\mathrm{OOD}} - \widehat{\Pi}_{\mathrm{OOD}}\|_F = 0$ and therefore

$$\mathcal{L}_{\mathrm{OOD}}(\widehat{\Pi}_{\mathrm{OOD}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}). \tag{24}$$

Applying the lower bound of Proposition 2 to $\Pi_{\mathrm{dec}} = \Pi_{\mathrm{ID}}$:

$$\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \;\geq\; \frac{\gamma_{\mathrm{OOD}}}{2}\,\Delta(\Pi_{\mathrm{ID}})^2. \tag{25}$$

Substituting $\mathcal{L}_{\mathrm{OOD}}(\widehat{\Pi}_{\mathrm{OOD}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})$ and rearranging:

$$\mathcal{L}_{\mathrm{OOD}}(\widehat{\Pi}_{\mathrm{OOD}}) = \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}}) \leq \mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{ID}}) - \frac{\gamma_{\mathrm{OOD}}}{2}\,\Delta(\Pi_{\mathrm{ID}})^2. \tag{26}$$

□

A.7 Proof of Corollary 1

Proof. Step 1 gives $(I - \widehat{\Pi}_{\mathrm{OOD}})\widetilde{W}_{\mathrm{dec}}(T^\star) = 0$. Left-multiplying Equation (8) by $I - \widehat{\Pi}_{\mathrm{OOD}}$ therefore gives

$$(I - \widehat{\Pi}_{\mathrm{OOD}})W_{\mathrm{dec}}^{\mathrm{GAE}} = R_\perp(B + \lambda_{\mathrm{geom}}I)^{-1}. \tag{27}$$

Since $B - \lambda_{\mathrm{pres}}I$ is a centered sample covariance, $B + \lambda_{\mathrm{geom}}I \succeq (\lambda_{\mathrm{pres}} + \lambda_{\mathrm{geom}})I$. Applying Equation (12) yields

$$\|(I - \widehat{\Pi}_{\mathrm{OOD}})W_{\mathrm{dec}}\|_F \leq \frac{(1-\alpha)\|R_\perp\|_F}{\lambda_{\mathrm{pres}} + \lambda_{\mathrm{geom}}}. \tag{28}$$

Let $U_r\Sigma_r V_r^\top$ be a top-r SVD truncation of W_{dec}, so $\Pi_{\mathrm{dec}} = U_r U_r^\top$. The matrix inequality $W_{\mathrm{dec}}^\top W_{\mathrm{dec}} \succeq W_{\mathrm{dec}}^\top \widehat{\Pi}_{\mathrm{OOD}} W_{\mathrm{dec}}$ implies $\sigma_r(W_{\mathrm{dec}}) \geq \sigma_r(\widehat{\Pi}_{\mathrm{OOD}}W_{\mathrm{dec}}) > 0$. Using $U_r = W_{\mathrm{dec}}V_r\Sigma_r^{-1}$, the equal-rank projector identity gives

$$\begin{aligned}
\|\Pi_{\mathrm{dec}} - \widehat{\Pi}_{\mathrm{OOD}}\|_F &= \sqrt{2}\|(I - \widehat{\Pi}_{\mathrm{OOD}})U_r\|_F \\
&\leq \frac{\sqrt{2}\|(I - \widehat{\Pi}_{\mathrm{OOD}})W_{\mathrm{dec}}\|_F}{\sigma_r(W_{\mathrm{dec}})} \\
&\leq \frac{2\|(I - \widehat{\Pi}_{\mathrm{OOD}})W_{\mathrm{dec}}\|_F}{\sigma_r(\widehat{\Pi}_{\mathrm{OOD}}W_{\mathrm{dec}})}.
\end{aligned} \tag{29}$$

The triangle inequality gives $\Delta(\Pi_{\mathrm{dec}}) \leq \varepsilon_N + \|\Pi_{\mathrm{dec}} - \widehat{\Pi}_{\mathrm{OOD}}\|_F$. Combining this with Equations (28) and (29), then applying the upper bound in Proposition 2, proves Equation (13).

At $\alpha = 1$, the rank condition on $W_{\mathrm{dec}}^{\mathrm{ID}}$ makes the Step 1-Only projector equal to $\widehat{\Pi}_{\mathrm{OOD}}$. Its faithfulness gap is ε_N, so Proposition 2 gives excess projection loss at most $\beta_{\mathrm{OOD}}\varepsilon_N^2/2$. This endpoint applies when Step 2 is skipped, including zero Step 2 regularization weights. □

A.8 Projection-Level Surrogate and Downstream Causal Metrics

This subsection relates projection loss to dictionary reconstruction and feature-intervention scores. The metrics are not equivalent to $\mathcal{L}_{\mathrm{OOD}}$. We bound activation-replacement loss with an additional reconstruction-error term and feature-score changes under stability assumptions.

Activation replacement. The dictionary reconstruction $\hat{h}(X) = W_{\mathrm{dec}}z(X) + b_{\mathrm{dec}}$ need not equal $\Pi_{\mathrm{dec}}h(X)$. Write $\Delta\mathrm{CE}(X) := \mathrm{CE}(\hat{h}(X)) - \mathrm{CE}(h(X))$, matching Equation (33).

Assumption 1 (Local smoothness)**.** *There exists $K_{\mathrm{CE}} < \infty$ such that $|\mathrm{CE}(\hat{h}(X)) - \mathrm{CE}(h(X))| \leq K_{\mathrm{CE}}\|h(X) - \hat{h}(X)\|_2$ for OOD activations and their dictionary reconstructions.*

Proposition 3 (Activation-replacement loss bound)**.** *Under Assumption 1, for $X \sim P_{\mathrm{OOD}}$,*

$$\mathbb{E}|\Delta\mathrm{CE}(X)| \le K_{\mathrm{CE}} \left(\sqrt{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})} + \sqrt{\mathbb{E}\|\Pi_{\mathrm{dec}} h(X) - \hat{h}(X)\|_2^2} \right).$$

The bound combines projection error with the additional error between the projection and the actual sparse reconstruction. Small projection error and small additional reconstruction error together yield a small upper bound on activation-replacement loss.

Proof.

$$\begin{aligned}
\mathbb{E}|\Delta\mathrm{CE}(X)| &\le K_{\mathrm{CE}}\, \mathbb{E}\|h - \hat{h}\|_2 \\
&\le K_{\mathrm{CE}} \left(\mathbb{E}\|h - \Pi_{\mathrm{dec}} h\|_2 + \mathbb{E}\|\Pi_{\mathrm{dec}} h - \hat{h}\|_2 \right) \\
&\le K_{\mathrm{CE}} \left(\sqrt{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{dec}})} + \sqrt{\mathbb{E}\|\Pi_{\mathrm{dec}} h - \hat{h}\|_2^2} \right). \qquad (30)
\end{aligned}$$

□

Feature interventions. Feature-intervention metrics require one more condition. Unlike ΔCE, nComp and nAOPC depend on which features are selected and ablated. Preserving feature activations does not ensure unchanged causal rankings after decoder adaptation. The following bound assumes stability of the selected feature sets and their intervention effects. Let $a_m := \frac{\ell_0 - \ell_{\setminus m}}{|\ell_0 - \ell_\varnothing|}$ denote the budget-m normalized logit drop in Equation (34) for the Fixed ID decoder $(W_{\mathrm{dec}}^{\mathrm{ID}}, b_{\mathrm{dec}}^{\mathrm{ID}})$, and let a_m^{GAE} denote the same quantity for the deployed GAE decoder $(W_{\mathrm{dec}}, b_{\mathrm{dec}})$. For the same masked feature vector v, the reconstruction difference is $\delta(v) := (W_{\mathrm{dec}} - W_{\mathrm{dec}}^{\mathrm{ID}})v + (b_{\mathrm{dec}} - b_{\mathrm{dec}}^{\mathrm{ID}})$. Write $z_{\setminus m}$ for z with the selected m feature activations set to zero.

Assumption 2 (Feature stability)**.** *For each feature budget used by nComp or nAOPC, both decoders use the same selected feature set and target token. The corresponding normalized logit-drop score is K_m-Lipschitz in the full, ablated, and empty replacement activations along this comparison, using the sum of their Euclidean distances. Both decoders use the same empty-reconstruction convention, either retaining their respective biases or setting the replacement activation to zero. The normalization denominators remain bounded away from zero.*

Proposition 4 (Feature-score stability under decoder adaptation)**.** *Under Assumption 2, for fixed decoders and $X \sim P_{\mathrm{OOD}}$,*

$$\mathbb{E}\left|a_m^{\mathrm{GAE}} - a_m\right| \le K_m \left(2\|W_{\mathrm{dec}} - W_{\mathrm{dec}}^{\mathrm{ID}}\|_F\, \mathbb{E}\|z\|_2 + 3\|b_{\mathrm{dec}} - b_{\mathrm{dec}}^{\mathrm{ID}}\|_2\right).$$

Averaging over budgets gives a bound for nAOPC with the average of K_m, and taking $m = m^*$ gives the nComp bound.

Proof. The full and ablated reconstruction changes are $\delta(z)$ and $\delta(z_{\setminus m})$. The empty-reconstruction change is $\delta(0)$ when biases are retained and zero under the common zero-activation convention. The Lipschitz assumption therefore gives

$$\begin{aligned}
\mathbb{E}\left|a_m^{\mathrm{GAE}} - a_m\right| &\le K_m\, \mathbb{E}\left(\|\delta(z)\|_2 + \|\delta(z_{\setminus m})\|_2 + \|\delta(0)\|_2\right) \\
&\le K_m \left(2\|W_{\mathrm{dec}} - W_{\mathrm{dec}}^{\mathrm{ID}}\|_F\, \mathbb{E}\|z\|_2 + 3\|b_{\mathrm{dec}} - b_{\mathrm{dec}}^{\mathrm{ID}}\|_2\right), \qquad (31)
\end{aligned}$$

The last step uses $\|z_{\setminus m}\|_2 \le \|z\|_2$ and the triangle inequality. Averaging the budget-wise bounds and applying the triangle inequality gives the nAOPC result. □

Proposition 4 controls score changes relative to Fixed under the stated assumptions. It gives no direction of improvement for nComp or nAOPC, which we evaluate empirically. Preserving feature activations does not establish the feature-stability condition.

B Empirical Evidence for Section 3

This appendix provides empirical support for the theoretical results in Section 3. We verify three claims: (i) the explainer subspace aligns with the ID-active subspace, (ii) the explainer-dependent component in the decomposition (3) accounts for a meaningful fraction of the total OOD error, and (iii) second-moment shift enlarges the faithfulness gap as predicted by Proposition 1.

B.1 Controlled Toy Setting

The toy experiments use a controlled setting that allows us to vary OOD severity continuously while keeping the model and explainer fixed.

Target model. The target model is a 2-layer ReLU MLP with input dimension d_{in}=128, hidden dimension d=256, and output dimension $p \in \{4, 8, 16\}$:

$$h(x) = \text{ReLU}(W_1 x + b_1) \in \mathbb{R}^d, \qquad o(x) = W_2\, h(x) + b_2 \in \mathbb{R}^p.$$

The hidden activations $h(x) \in \mathbb{R}^d$ correspond to the hidden representations analyzed in the main text. The explainer operates on these d-dimensional activations.

Explainer. We train both a transcoder and an SAE on ID hidden activations using the standard reconstruction-plus-sparsity objective (ERM), with dictionary sizes $k \in \{d/2,\ 1d,\ 2d,\ 4d,\ 8d,\ 32d\}$. The subspace rank is set to $r = p$, matching the rank of the output weight matrix $W_2 \in \mathbb{R}^{p\times d}$: since $o(x) = W_2\, h(x) + b_2$, the model's output depends on $h(x)$ only through its projection onto $\text{span}(W_2^\top)$, which has dimension p. We use $r = p$ as a truncation choice for the hidden second moment. Its leading eigenspace need not coincide with the output-relevant subspace $\text{span}(W_2^\top)$.

OOD generation. ID inputs are drawn from $x \sim \mathcal{N}(0, I_{d_{\text{in}}})$. OOD inputs are generated as $x = A_s\, z$ where $z \sim \mathcal{N}(0, I_{d_{\text{in}}})$ and A_s is a severity-dependent transformation matrix. The resulting input covariance is $\Sigma(s) = A_s A_s^\top$, which is symmetric and positive semidefinite. The covariance is globally normalized so that $\text{tr}(\Sigma(s)) = d_{\text{in}}$ for all s, ensuring that reconstruction error differences are not driven by trivial scale changes. At s=0, $\Sigma(0) = I$ (ID). As s increases toward 1, the input covariance undergoes progressive rotation and anisotropic rescaling, which propagates through the ReLU layer to induce second-moment shift in hidden space. We use N=20,000 samples for all geometric computations.

Validation on real models. We also validate the results on pretrained language models (GPT-2 Small, Pythia-1.4B) under temporal, domain, and adversarial distribution shifts, using the experimental setup described in Section 5. Unlike the toy setting, OOD severity is not varied continuously; instead, we compare ID and pure OOD for each shift type.

B.2 Explainer Subspace Alignment with the ID Dominant Subspace

Section 3.1 claims that the explainer subspace Π aligns closely with Π_{ID} for a well-trained ID explainer. We verify this by measuring the subspace overlap

$$\text{overlap}(U_{\text{dec}},\, U_{\text{ID}}) := \frac{1}{r}\, \|U_{\text{dec}}^\top U_{\text{ID}}\|_F^2 \ \in\ [0, 1],$$

where $U_{\text{ID}} \in \mathbb{R}^{d\times r}$ contains the top-r eigenvectors of M_{ID}. A value of 1 indicates perfect alignment; 0 indicates orthogonality.

B.2.1 Toy Setting

We sweep dictionary sizes $k \in \{d/2,\ 1d,\ 2d,\ 4d,\ 8d,\ 32d\}$ for both transcoders and SAEs to test whether the claim holds across different levels of overcompleteness. Figure 5 reports the overlap as a function of OOD severity for each dictionary size. In both panels, solid lines show the explainer–ID overlap and dashed lines show the explainer–OOD overlap.

Transcoder (left). The explainer–ID overlap (solid) remains above 0.89 for all dictionary sizes and all severity levels. The overlap is largely insensitive to k: even an undercomplete dictionary (k=$d/2$) captures the ID-active subspace well. The explainer–OOD overlap (dashed) degrades monotonically with severity, dropping below 0.3 at s=1.0 regardless of k.

SAE (right). The explainer–ID overlap depends more on k. For $k \geq 4d$, the overlap exceeds 0.93, comparable to the transcoder. For smaller k (k=$d/2$ or k=$1d$), the overlap drops to 0.70–0.77, indicating that undercomplete SAEs do not fully capture the ID-active subspace. As with the transcoder, the explainer–OOD overlap degrades with severity for all k.

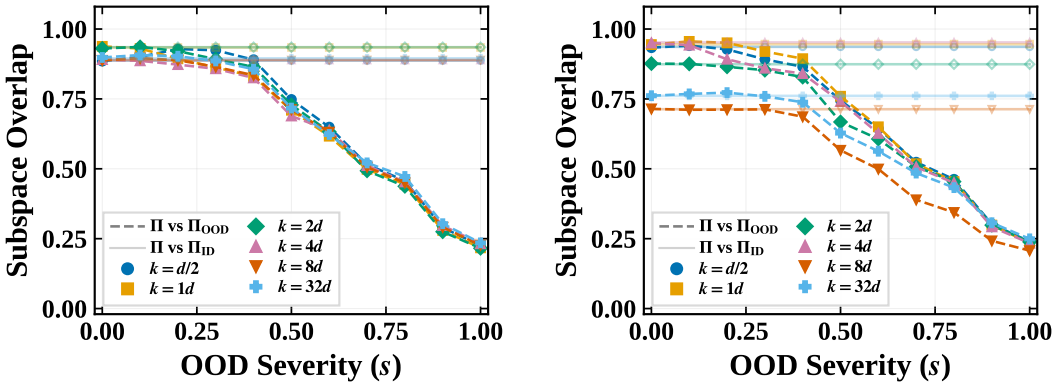


Figure 5: **Explainer subspace overlap between the explainer and the ID-active subspace (solid) vs. the OOD-active subspace (dashed), as a function of OOD severity s, for dictionary sizes $k \in \{d/2, 1d, 2d, 4d, 8d, 32d\}$. Left:** Transcoder. **Right:** SAE. For $k \geq 4d$, both explainer types maintain high ID overlap (> 0.89) regardless of severity, while OOD overlap degrades monotonically.

Interpretation. For sufficiently overcomplete dictionaries ($k \geq 4d$, the standard setting in practice), both transcoders and SAEs align closely with the ID-active subspace regardless of OOD severity, confirming the claim in Section 3.1. The divergence between the ID overlap (flat) and the OOD overlap (decreasing) is precisely the faithfulness gap: the explainer remains anchored to the ID geometry while the model's active subspace rotates away under OOD shift.

B.2.2 Real-Data Setting

We evaluate ID-trained explainers on GPT-2 Small and Pythia-1.4B under temporal, domain, and adversarial shifts. The rank is r=64 for both models.

Faithfulness gap on M_{ID} vs. on M_{OOD}. Definition 1 measures the gap of Π_{dec} against Π_{OOD}. To assess the residual ID-side misalignment, we additionally report the analogous quantity against Π_{ID},

$$\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}}) := \|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{dec}}\|_F,$$

which directly quantifies how well the ID-trained explainer captures Π_{ID}. Frobenius gap and overlap encode the same information through $\|\Pi_A - \Pi_B\|_F = \sqrt{2r\,(1 - \mathrm{overlap}(U_A, U_B))}$, and Table 6 reports the corresponding gaps. The maximum value at r=64 is $\sqrt{2r} = \sqrt{128} \approx 11.31$.

Table 6: **Frobenius faithfulness gap of the ID-trained explainer at r=64.** Columns 4–6 report $\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}}) = \|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{dec}}\|_F$, $\Delta(\Pi_{\mathrm{dec}}) = \|\Pi_{\mathrm{OOD}} - \Pi_{\mathrm{dec}}\|_F$, and $\|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F$. The last column is the ratio $\Delta(\Pi_{\mathrm{dec}})/\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$. Maximum possible value at r=64 is $\sqrt{128} \approx 11.31$. All values are derived from the measured subspace overlaps.

Explainer	Model	Shift	$\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$	$\Delta(\Pi_{\mathrm{dec}})$	$\|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F$	Ratio
Transcoder	GPT-2 Small	Temporal	5.89	6.17	5.65	1.05×
	GPT-2 Small	Domain	5.89	9.79	9.83	1.66×
	GPT-2 Small	Adversarial	5.89	9.97	9.83	1.69×
	Pythia-1.4B	Temporal	7.73	8.72	7.68	1.13×
	Pythia-1.4B	Domain	7.73	10.75	10.78	1.39×
	Pythia-1.4B	Adversarial	7.73	10.46	10.49	1.35×
SAE	GPT-2 Small	Temporal	9.76	9.75	4.83	1.00×
	GPT-2 Small	Domain	9.76	10.98	10.33	1.12×
	GPT-2 Small	Adversarial	9.76	10.95	10.30	1.12×
	Pythia-1.4B	Temporal	9.95	10.20	7.65	1.03×
	Pythia-1.4B	Domain	9.95	11.10	10.88	1.12×
	Pythia-1.4B	Adversarial	9.95	11.03	10.71	1.11×

For transcoders, $\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$ is consistently smaller than $\Delta(\Pi_{\mathrm{dec}})$, with the OOD-side gap exceeding the ID-side gap by 1.35–1.69$\times$ under domain and adversarial shifts, where the second-moment shift is largest. Moreover, $\|\Pi_{\mathrm{ID}} - \Pi_{\mathrm{OOD}}\|_F$ tracks $\Delta(\Pi_{\mathrm{dec}})$ to within 3% under these shifts. The residual ID-side misalignment still enters the triangle-inequality bound in Section 3.2. SAE explainers exhibit a larger $\Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$ (close to the upper bound at r=64), reflecting the lower per-rank overlap of residual-stream dictionaries; the ordering $\Delta(\Pi_{\mathrm{dec}}) \geq \Delta_{\mathrm{ID}}(\Pi_{\mathrm{dec}})$ still holds under domain and adversarial shifts but with a smaller margin (1.11–1.12$\times$). Temporal shift produces nearly equal ID- and OOD-side gaps for both explainer types, consistent with its milder second-moment perturbation.

B.3 Relative Magnitude of the Explainer-Dependent Term

The decomposition of Equation (3) separates OOD faithfulness as

$$\mathcal{L}_{\mathrm{OOD}}(\Pi) = \underbrace{\mathcal{L}_{\mathrm{OOD}}(\Pi_{\mathrm{OOD}})}_{\text{irreducible}} + \underbrace{\mathrm{tr}\big((\Pi_{\mathrm{OOD}} - \Pi)\, M_{\mathrm{OOD}}\big)}_{\text{explainer-dependent}}.$$

Adaptation can only reduce the explainer-dependent component. If this component is negligible relative to the irreducible component, no adaptation strategy can meaningfully improve OOD faithfulness. We measure the explainer-dependent ratio

$$\eta := \frac{\mathrm{tr}[(\Pi_{\mathrm{OOD}} - \Pi)\, M_{\mathrm{OOD}}]}{\mathcal{L}_{\mathrm{OOD}}(\Pi)}$$

to assess whether the explainer-dependent component is an actionable target.

B.3.1 Toy Setting

In the toy setting, η rises from below 0.05 for most k at $s = 0$ to approximately 0.31 at $s = 1$, with little variation across dictionary sizes. The moderate value reflects the low rank ratio $r/d = 8/256$, which leaves a large irreducible component in this setting.

B.3.2 Real-Data Setting

Figure 6 reports η at pure ID and pure OOD for GPT-2 Small and Pythia-1.4B under temporal, domain, and adversarial shifts.

Transcoder (left). At pure OOD, domain and adversarial shifts reach $\eta > 0.99$ across both models: the explainer-dependent component dominates the total error almost entirely. Temporal shift yields $\eta \approx 0.66$–0.99 depending on the model, reflecting its milder geometric distortion. These values are substantially higher than in the toy setting ($\eta \approx 0.31$), because the toy setting uses $r/d = 8/256 = 3.1\%$ so the irreducible term dominates by construction.

SAE (right). SAE explainers exhibit the same trend. Under domain and adversarial shifts, $\eta > 0.99$ for both models, confirming that the explainer-dependent component dominates. Temporal shift yields $\eta \approx 0.68$–0.99 depending on the model, consistent with its milder geometric distortion. At pure ID, $\eta \approx 0.51$ (GPT-2) and $\eta \approx 0.99$ (Pythia-1.4B), reflecting the larger model's higher effective dimensionality relative to r=64.

B.4 Empirical Verification of Proposition 1

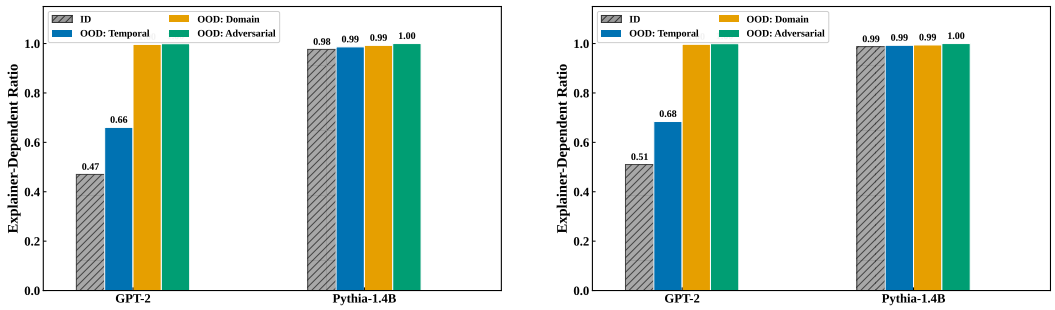
Proposition 1 predicts that second-moment shift controls the faithfulness gap via

$$\Delta(\Pi_{\mathrm{ID}}) \leq \frac{2}{\gamma_{\mathrm{ID}}} \|M_{\mathrm{OOD}} - M_{\mathrm{ID}}\|_F.$$

Since this bound depends only on M_{ID} and M_{OOD}, it is independent of the explainer architecture and dictionary size. We verify it empirically on both the toy setting and real models.

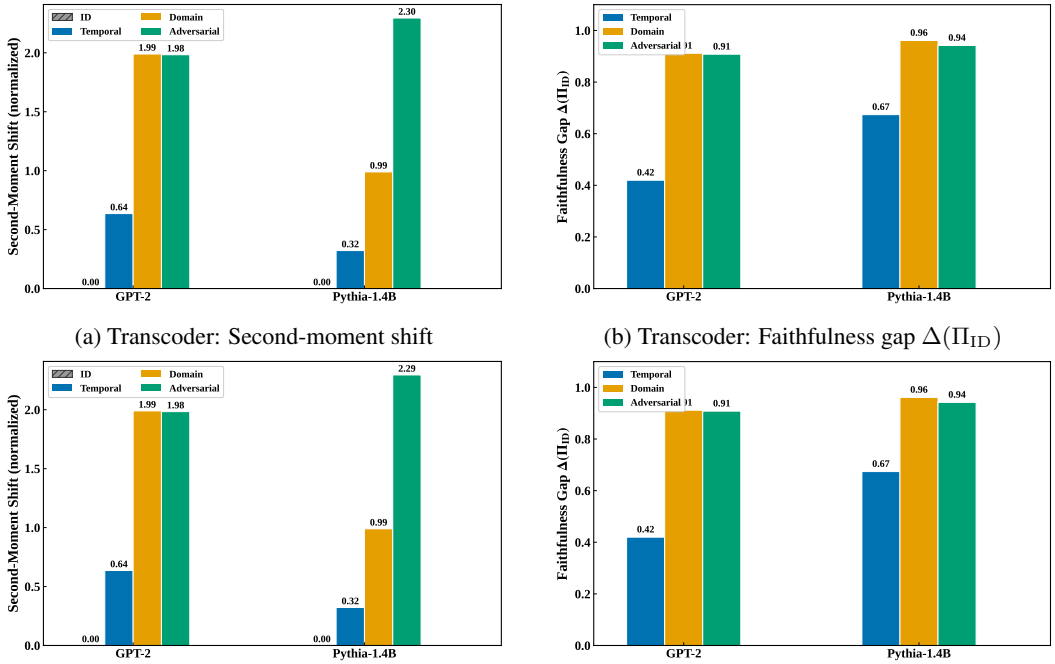
B.4.1 Toy Setting

In the toy setting, normalized second-moment shift and $\Delta(\Pi_{\mathrm{ID}})$ are strongly correlated (Pearson $r = 0.993$, Spearman $\rho = 1.000$).



(a) Transcoder (b) SAE

Figure 6: **Explainer-dependent ratio η at pure ID (hatched) and pure OOD (colored) (r=64).** Under domain and adversarial shifts, $\eta > 0.99$ for both explainer types.



(a) Transcoder: Second-moment shift (b) Transcoder: Faithfulness gap $\Delta(\Pi_{\mathrm{ID}})$

(c) SAE: Second-moment shift (d) SAE: Faithfulness gap $\Delta(\Pi_{\mathrm{ID}})$

Figure 7: **Proposition 1 verification (r=64) at pure OOD.** Top row: Transcoder. Bottom row: SAE. Within each model, larger shifts correspond to larger gaps for both explainer types.

B.4.2 Real-Data Setting

Figure 7 shows the second-moment shift and faithfulness gap for GPT-2 Small and Pythia-1.4B (r=64) at pure OOD for both transcoders (top row) and SAEs (bottom row). Within each model, the ordering is consistent: temporal shift produces the smallest second-moment shift and the smallest faithfulness gap, while domain and adversarial shifts produce larger shifts and correspondingly larger gaps.

SAE. The same ordering holds for SAE explainers: temporal shift produces the smallest second-moment shift and faithfulness gap, while domain and adversarial shifts produce larger values. The magnitudes are comparable to transcoders, confirming that the proposition is independent of the explainer architecture.

Summary. The proposition holds across all models and both explainer types under diverse real-world distribution shifts, confirming that the theoretical predictions generalize beyond the controlled toy setting.

C Empirical Illustration of Theorem 1

Under exact subspace estimation, Theorem 1 lower-bounds the Step 1 projection-loss improvement by $\frac{1}{2}\gamma_{\mathrm{OOD}}\,\Delta(\Pi_{\mathrm{ID}})^2$. We illustrate this relationship in the controlled toy setting (Section 5.1) by fixing $r = p = 8$ and sweeping OOD severity from s=0 (ID) to s=1 (maximum shift). At each severity, we use the Step 1 projector $\widehat{\Pi}_{\mathrm{OOD}}$ and measure the projection-loss improvement $I(s) = \mathcal{L}_{\mathrm{OOD},s}(\Pi_{\mathrm{ID}}) - \mathcal{L}_{\mathrm{OOD},s}(\widehat{\Pi}_{\mathrm{OOD}})$ against the squared faithfulness gap $\Delta(\Pi_{\mathrm{ID}})^2$.

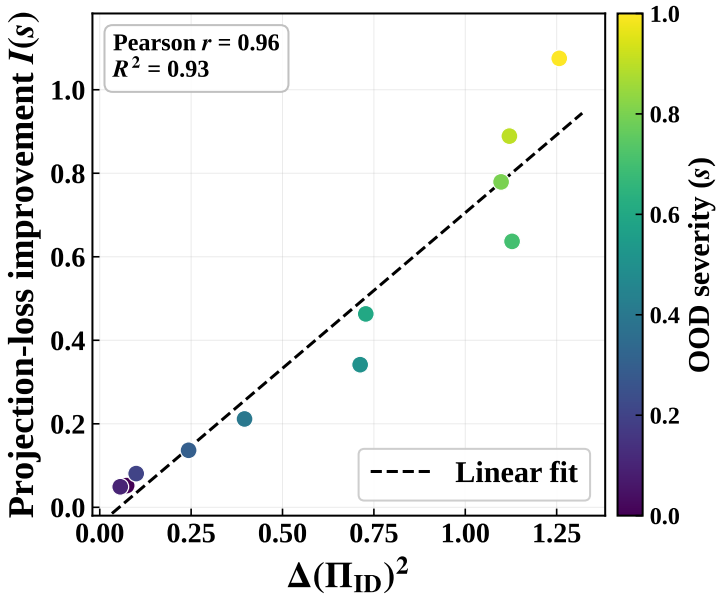


Figure 8: **Empirical illustration of Theorem 1 on the controlled toy setting.** Projection-loss improvement $I(s)$ versus the squared faithfulness gap $\Delta(\Pi_{\mathrm{ID}})^2$, swept across OOD severity $s \in [0, 1]$. The dashed line is a linear fit ($R^2 = 0.93$, Pearson $r = 0.96$), illustrating the Step 1 relationship.

Figure 8 shows a strong linear relationship between $I(s)$ and $\Delta(\Pi_{\mathrm{ID}})^2$ (Pearson $r = 0.96$, $R^2 = 0.93$), illustrating the Step 1 relationship. The empirical improvement exceeds the population lower-bound expression at every severity (0 violations out of 11). This finite-sample comparison illustrates Step 1 behavior under estimated OOD subspaces. The reference curve uses the eigengap coefficient $\gamma_{\mathrm{OOD}}/2$, which is conservative relative to the effective improvement rate, as expected for a spectral-gap-based guarantee.

D Experimental Details

D.1 Model and Explainer Details

Table 7 summarizes the model and explainer configurations. All models are frozen pretrained checkpoints; only explainer components are adapted.

For each model, we train two explainer types: Top-K SAEs [7], which reconstruct residual-stream activations h using Top-K sparsity, and transcoders [5], which reconstruct MLP outputs h from sparse feature activations z extracted from MLP inputs. All explainers use dictionary size k=32d [4] and are trained on in-distribution activations with the standard reconstruction-plus-sparsity objective.

Table 7: **Model and explainer configurations.**

Model	d	Layer	k ($32d$)	ID Corpus
GPT-2 Small	768	8	24,576	OpenWebText [47]
Pythia-1.4B	2,048	15	65,536	The Pile [48]

D.2 Baseline Details

- **TERM.** Tilted empirical risk minimization uses only ID data [16, 17].
- **Retrain.** The explainer retrained from scratch on OOD activations with the same architecture and hyperparameters. This baseline provides a reference point but is not an oracle upper bound: retraining on OOD data can distort pretrained feature structure [49].
- **SAEBoost.** A residual boosting approach [18]: a secondary explainer is trained on the OOD reconstruction residuals of the ID-trained base explainer, and the two outputs are summed at inference ($\hat{h} = \hat{h}_{\text{base}} + \hat{h}_{\text{resid}}$). This adds OOD-specific capacity while retaining the base dictionary, but requires OOD training data.
- **FaithfulSAE.** Full retraining uses only the model's own unconditional generations [15].

D.3 Comparison with Geometric Domain Adaptation

The geometric baselines use the same unlabeled OOD activation cache and evaluation protocol as GAE. Table 4 reports the HaluEval transcoder results in the main text.

Baseline definitions. PCA estimates the rank-r OOD principal subspace and uses its signed coefficients for evaluation. Its dense coordinates differ from sparse dictionary features in sparsity and feature count, so this comparison is approximate. For SA [30] and CORAL [31], we keep the sparse encoder frozen and transform only the decoder. These baselines are decoder adaptations of the original domain adaptation methods.

Additional FineWeb evaluation. Table 8 compares the same methods on the Pythia-1.4B SAE under FineWeb. GAE achieves higher nAOPC and nComp and lower $|\Delta\text{CE}|$ than PCA, SA, and CORAL in this setting.

Table 8: **Geometric adaptation on FineWeb (Pythia-1.4B, SAE).** GAE improves both causal faithfulness and reconstruction quality over PCA, SA, and CORAL. **Bold**: best. Underline: second best.

Method	**nAOPC↑**	**nComp↑**	**$\lvert\Delta$CE$\rvert$↓**
Fixed	<u>0.962</u>	<u>1.536</u>	<u>0.0216</u>
PCA	0.959	0.978	0.1324
SA	0.841	1.090	0.1301
CORAL	0.953	1.247	0.1722
GAE	**0.985**	**1.677**	**0.0207**

D.4 GAE Implementation Details

Step 2 dictionary preservation. The dictionary preservation penalty in the closed-form decoder refit (Section 4.2, Step 2) regularizes the decoder toward the Step 1 output $\widetilde{W}_{\text{dec}}(T^\star)$ with weight λ_{pres}, following Equation (8).

Decoder interpolation. The Step 2 closed-form solution $W_{\text{dec}}^{\text{GAE}}$ optimizes sample-level reconstruction under geometry penalties. With limited OOD samples, this solution can overfit to the estimation noise in $\{z_i, h_i\}$. We use the decoder interpolation in Equation (12) to limit this effect. We treat α as a hyperparameter selected per OOD setting.

Hyperparameter selection. GAE requires no gradient computation or iterative optimization. The hyperparameters (r, λ_{geom}, λ_{pres}, α) are selected per OOD setting using a small held-out portion of unlabeled OOD activations, monitoring reconstruction quality ($|\Delta\text{CE}|$) and causal faithfulness (nComp). We sample the OOD validation/adaptation split with a fixed seed from the same OOD source. It contains N_{cov}=2,000 prompts for estimating $\widehat{M}_{\text{OOD}}$ and N_{fit} activations for Step 2. No OOD labels are required. The final test split contains N_{eval}=1,000 *disjoint* prompts. Once selected, the same hyperparameters are used for all evaluation prompts.

Hyperparameter summary. Tables 9 and 10 list the GAE hyperparameters for transcoders and SAEs, respectively. The search grid uses $r \in \{4, 8, 32, 64, 256, 512, 1024\}$ for transcoders and explainer-specific SAE rank grids, $\lambda_{\text{geom}} \in \{0, 0.1, 1, 15, 20\}$, $\lambda_{\text{pres}} \in \{0, 0.04, 0.2, 1, 2\}$, $\alpha \in \{0, 1\}$, and $N_{\text{fit}} \in \{0, 2048\}$. All settings use the second-moment matrix (not centered covariance) for OOD subspace estimation, as prescribed in Algorithm 1.

Table 9: **GAE hyperparameters for transcoder experiments.** N_{fit} counts hidden-state activation vectors used for Step 2.

Model	OOD Setting	r	λ_{geom}	λ_{pres}	α	N_{fit}
GPT-2	HaluEval (Adversarial)	32	0.1	0.2	0	2,048
GPT-2	Edgar (Domain)	4	0.1	0.2	0	2,048
GPT-2	FineWeb (Temporal)	8	0.1	0.04	0	2,048
Pythia-1.4B	HaluEval (Adversarial)	64	15	2	0	2,048
Pythia-1.4B	Edgar (Domain)	64	0.1	0.2	0	2,048
Pythia-1.4B	FineWeb (Temporal)	64	20	1	0	2,048

Table 10: **GAE hyperparameters for SAE experiments.** N_{fit} counts hidden-state activation vectors used for Step 2. In Step 1-Only (Procrustes rotation), decoder refitting is skipped, $N_{\text{fit}} = 0$, and $\alpha = 1$. The unused Step 2 regularization weights are listed as zero.

Model	OOD Setting	r	λ_{geom}	λ_{pres}	α	N_{fit}
GPT-2	HaluEval (Adversarial)	512	0	0	1	0
GPT-2	Edgar (Domain)	512	0	0	1	0
GPT-2	FineWeb (Temporal)	512	0	0	1	0
Pythia-1.4B	HaluEval (Adversarial)	4	0.1	0.2	0	2,048
Pythia-1.4B	Edgar (Domain)	1,024	0	0	1	0
Pythia-1.4B	FineWeb (Temporal)	4	0	0.2	0	2,048

D.5 Evaluation Details

Normalized comprehensiveness (nComp). We measure causal faithfulness via logit-level feature ablation [4, 21]. Given a prompt, let ℓ_0 denote the target-token logit under the explainer's full reconstruction, and $\ell_\varnothing$ the logit when all features are ablated to zero. For a feature budget m^*, let $\ell_{\setminus m^*}$ be the logit after removing the top-m^* features. We define

$$\text{nComp} = \frac{\ell_0 - \ell_{\setminus m^*}}{|\ell_0 - \ell_\varnothing|}, \tag{32}$$

where $m^* = 32$. Higher values indicate that the top features are causally important for the model's output.

Delta cross-entropy (ΔCE). We measure reconstruction quality by the cross-entropy change when original activations are replaced with the explainer's reconstruction [7]:

$$\Delta\text{CE} = \text{CE}(\hat{h}) - \text{CE}(h), \tag{33}$$

where $\text{CE}(h)$ is the loss with original activations and $\text{CE}(\hat{h})$ is the loss with reconstructed activations. We report $|\Delta\text{CE}|$, with values closer to zero indicating better reconstruction quality.

Normalized AOPC (nAOPC). nAOPC [23] averages the normalized logit drop across multiple feature budgets when top-m features are removed:

$$\mathrm{nAOPC} = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \frac{\ell_0 - \ell_{\setminus m}}{|\ell_0 - \ell_\varnothing|}, \tag{34}$$

where $\mathcal{M} = \{1, 2, 4, 8, 16, 32, 64, 128\}$. Higher values indicate that the identified features are causally important across a range of budgets.

Evaluation protocol. We evaluate at the last token position using zero-residual ablation (replacing ablated features with zero). The denominator $|\ell_0 - \ell_\varnothing|$ normalizes each example by the logit range between full and empty reconstruction, enabling cross-example comparison regardless of the absolute logit scale. We exclude examples where $|\ell_0 - \ell_\varnothing| < 0.1$ to avoid unstable normalization. Feature budgets are $\mathcal{M} = \{1, 2, 4, 8, 16, 32, 64, 128\}$ with $m^* = 32$. We use $N_{\mathrm{eval}} = 1{,}000$ evaluation prompts per setting and seed $= 2026$ throughout.

D.6 Compute Resources

All experiments run on a single NVIDIA RTX A6000 GPU (48 GiB VRAM) with an Intel Xeon Gold 6326 CPU and 252 GiB of system RAM. No experiment requires multi-GPU or model-parallel execution. GPT-2 runs use peak GPU memory under 8 GiB. Pythia-1.4B runs with batch size 64 use peak GPU memory under 24 GiB.

Per-method wall-clock. Table 1 reports the cost of a single (model, OOD setting) run for each adaptation method. GAE times are measured with fixed hyperparameters and exclude hyperparameter search. GAE uses a single forward pass over ∼2,000 OOD activations and no gradient computation. Faithfulness evaluation (nAOPC, nComp, ΔCE on 1,000 prompts) adds about 1 minute per (model, OOD setting, baseline).

Total compute. The full result table (two models, three OOD settings, six adaptation baselines) requires roughly 50 GPU-hours on a single RTX A6000, dominated by the Retrain-style baselines on Pythia-1.4B. The GAE rows themselves contribute under 1 GPU-minute to this total. Pretraining of the ID dictionaries (transcoders and SAEs on OpenWebText / The Pile) is a one-time cost that we treat as external to the adaptation experiments.

E Additional Case Study and Concept-Level Retention

This appendix gives the protocol for Section 5.2.3 and applies it to a HaluEval prompt predicting a male first name. We keep the encoder frozen and select the top-3 features by GAE causal effect on the target token. Fixed and GAE use identical feature IDs and activations for each prompt.

Each feature's direct logit attribution (DLA) is its activation multiplied by the dot product of its decoder column and the candidate token's unembedding vector. For each feature, we subtract its mean DLA on 10 unrelated noun controls from its mean DLA on 20 class-member tokens. The total class-specificity score sums these differences over the three selected features.

For this prompt, the same three features with unchanged activations contribute more to the first-name class under GAE than under Fixed.

Concept-level retention. The preceding case study selects features by their causal effect under GAE. To assess retention of the original concept relationships, we select three features per concept family using only the Fixed explainer on a disjoint ID split. We keep these feature IDs fixed across Fixed, Step 1-Only, and GAE, and evaluate them on unseen HaluEval prompts. Feature selection uses neither GAE scores nor OOD evaluation prompts.

Table 11 shows that GAE retains the Fixed DLA sign on 92.8–96.1% of feature–prompt pairs, with Fixed–GAE DLA correlations of 0.86–0.93. Mean class-specific DLA increases for all three concept families, and all paired-bootstrap confidence intervals exclude zero. These results support concept-selective retention for the evaluated features and prompts, without establishing universal semantic or causal invariance.

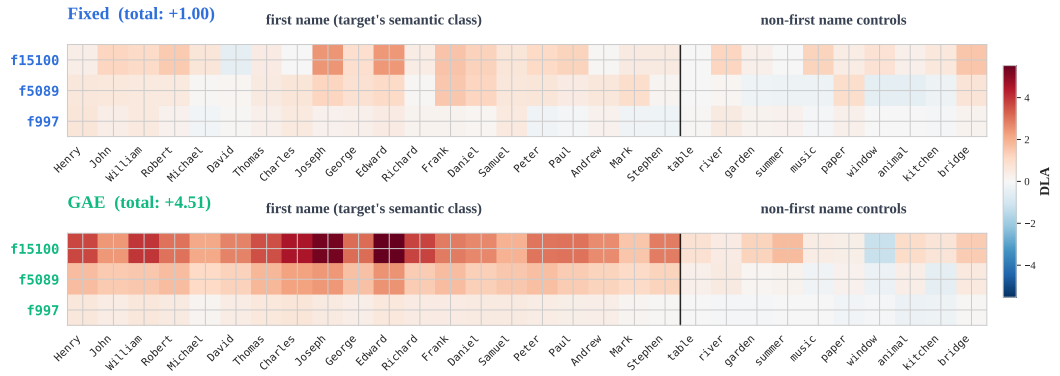


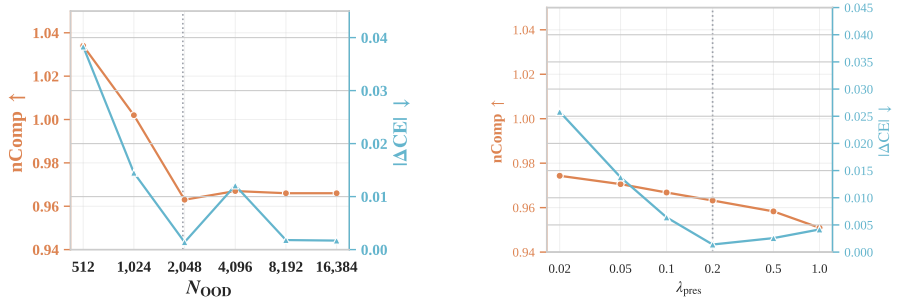
Figure 9: **Per-feature DLA on a prompt predicting ' Henry' (GPT-2, Transcoder).** The truncated input is "Question: What nationality was James"; the next token is a male first name. Each cell reports the feature's direct logit attribution to a candidate token, with 20 male first names on the left and 10 unrelated noun controls on the right. Fixed's total class-specificity is $+1.00$ and GAE's is $+4.51$, a $4.5\times$ amplification of the same selected features' pull toward the first-name class.

Table 11: **Concept-level retention on unseen HaluEval prompts.** Fixed, Step 1-Only, and GAE columns report mean class-specific DLA for the same ID-selected features. Sign retention and correlation compare Fixed with GAE. The 95% confidence intervals are paired-bootstrap intervals for the mean GAE-minus-Fixed DLA difference.

Concept	Mean DLA ↑			Sign retention	Corr.	95% CI
	Fixed	Step 1-Only	GAE			
Nationality	1.31	1.24	**1.39**	94.4%	0.89	[0.03, 0.13]
First name	1.68	1.57	**1.72**	96.1%	0.93	[0.01, 0.08]
Profession	1.18	1.10	**1.27**	92.8%	0.86	[0.04, 0.15]

F Hyperparameter Sensitivity

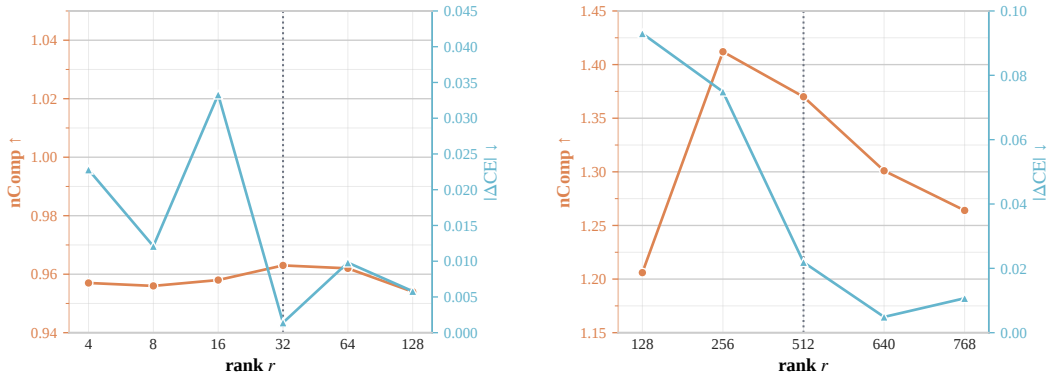
We examine GAE's sensitivity to OOD sample size N_{OOD} and dictionary preservation weight λ_{pres} on HaluEval with GPT-2 transcoders (Figure 10). We additionally compare rank sensitivity for GPT-2 transcoders and Top-K SAEs (Figure 11).



(a) N_{OOD}. (b) λ_{pres}.

Figure 10: **Hyperparameter sweeps on HaluEval (GPT-2, Transcoder).** nComp (orange, left axis) and $|\Delta\text{CE}|$ (cyan, right axis) respond differently to the adaptation settings, motivating joint validation of causal faithfulness and reconstruction quality.

Rank r: For GPT-2 transcoders on HaluEval, nComp varies little across the evaluated ranks, while reconstruction quality is more sensitive. At $r = 32$, nComp is 0.963 and $|\Delta\text{CE}|$ is 0.0014; extending



(a) Transcoder.

(b) Top-K SAE.

Figure 11: **Rank sensitivity on HaluEval (GPT-2).** Transcoders and Top-K SAEs exhibit different rank-dependent trade-offs, supporting explainer-specific validation rather than a universal rank choice. Orange circles show nComp (left axis), and cyan triangles show $|\Delta\mathrm{CE}|$ (right axis), with axis ranges differing between panels. Dotted lines mark the ranks used in the main experiments.

the sweep to $r = 128$ gives 0.954 and 0.0058, respectively. The Top-K SAE sweep shows a different trade-off: $r = 256$ gives the highest nComp (1.412), whereas $r = 640$ gives the lowest $|\Delta\mathrm{CE}|$ (0.0049). Thus, r controls the retained OOD geometry rather than representing a universal intrinsic dimension of distribution shift. We select rank on a validation split by jointly monitoring nComp and $|\Delta\mathrm{CE}|$, reusing the same OOD activations and eigenvectors across ranks.

OOD sample size N_{OOD}: Increasing the number of OOD activation vectors does not monotonically improve the downstream metrics. From $N_{\mathrm{OOD}} = 2{,}048$ to $4{,}096$, nComp changes little (0.963 to 0.967), but $|\Delta\mathrm{CE}|$ increases from 0.0014 to 0.0121. At $N_{\mathrm{OOD}} = 8{,}192$ and $16{,}384$, $|\Delta\mathrm{CE}|$ returns to 0.0018 and 0.0017, respectively, while nComp is 0.966 at both sample sizes. Although $N_{\mathrm{OOD}} = 512$ gives the highest nComp (1.034), it also gives the largest reconstruction drift (0.0383). These results support $N_{\mathrm{OOD}} = 2{,}048$ as a favorable balance in this setting, rather than a universally optimal sample size.

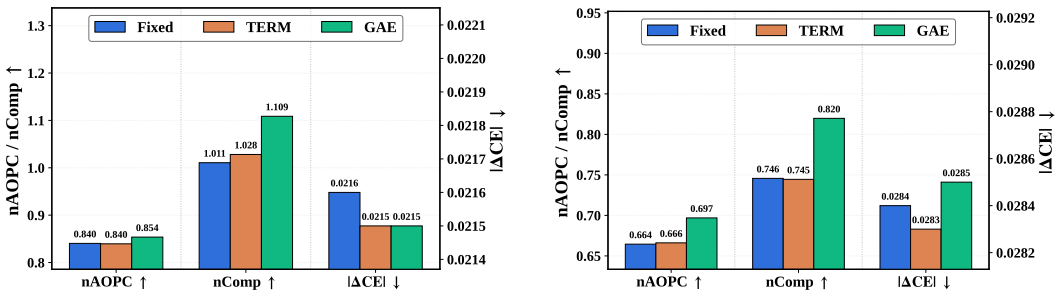
Dictionary preservation weight λ_{pres}: increasing λ_{pres} trades a small nComp decrease (0.02) for a large $|\Delta\mathrm{CE}|$ improvement ($0.026 \rightarrow 0.001$), with the default near the elbow.

G Faithfulness on Held-out In-Distribution Data

We evaluate GAE against the two training-free baselines (Fixed, TERM) on a held-out subset of each model's training corpus (OpenWebText for GPT-2, the Pile for Pythia-1.4B). Training-based baselines (Retrain, Finetune, SAEBoost, FaithfulSAE) are outside this comparison because they target a different operating regime, exploiting sample-level specialization rather than geometric adaptation, so mixing them would confound the cause of any improvement.

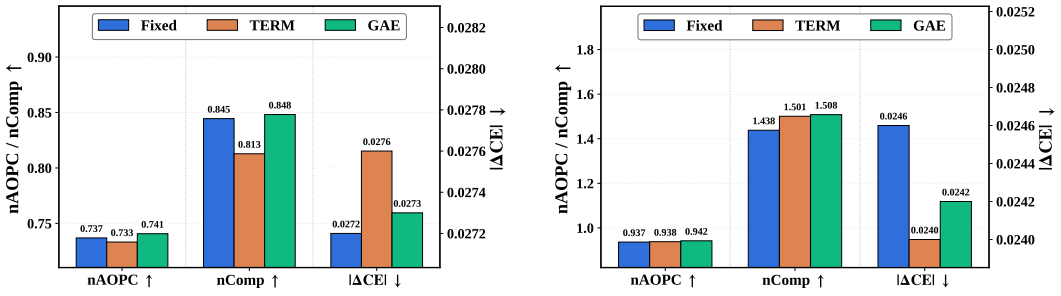
Each model is paired with a held-out slice of its own training corpus that no explainer has touched during dictionary fitting. The evaluation prompts and the 2,000-sample adaptation set are drawn fresh from this slice, sharing the same broad domain as the data the Fixed explainer was trained on while remaining strictly unseen. The resulting setup isolates the no-shift regime at the distribution level: there is no domain gap and no temporal gap, only the sampling variation that any finite subset inherits from a large corpus. The question this section asks is whether a geometric adaptation method has anything to do once the obvious shift has been removed.

The answer is yes. Even on this ID slice, the 2,000 adaptation samples carry their own subset-specific second-moment structure, which is statistically distinct from the full training corpus the Fixed checkpoint saw. GAE picks up this finer geometry: it improves both causal metrics over Fixed in every cell, while leaving reconstruction quality nearly unchanged. The Transcoder cells show the largest correction, with nComp moving from 1.011 to 1.109 on GPT-2 and from 0.746 to 0.820 on Pythia-1.4B, and nAOPC tracking ($0.840 \rightarrow 0.854$ and $0.664 \rightarrow 0.697$). The Top-K SAE cells



(a) GPT-2 (b) Pythia-1.4B

Figure 12: **Faithfulness of training-free explainer methods on held-out in-distribution data with the Transcoder explainer.** The left axis plots the causal-faithfulness metrics nAOPC and nComp (higher is better) and the right axis plots reconstruction quality $|\Delta\mathrm{CE}|$ (lower is better). Both backbones show the same pattern: GAE lifts nAOPC and nComp above Fixed and TERM, with the largest swing on GPT-2 (nComp $+0.10$), while $|\Delta\mathrm{CE}|$ tracks Fixed to within 0.0001.



(a) GPT-2 (b) Pythia-1.4B

Figure 13: **Faithfulness of training-free explainer methods on held-out in-distribution data with the Top-K SAE explainer.** The left axis plots nAOPC and nComp (higher is better) and the right axis plots $|\Delta\mathrm{CE}|$ (lower is better). The SAE dictionary already sits closer to optimal on this ID slice, so the gap to Fixed is narrower than on the Transcoder cells, but GAE still moves both causal metrics in the right direction (largest at Pythia-1.4B with nComp $+0.07$), and on Pythia-1.4B even nudges $|\Delta\mathrm{CE}|$ slightly below Fixed ($0.0246 \rightarrow 0.0242$).

move less but in the same direction, with the largest jump at Pythia-1.4B where nComp rises from 1.438 to 1.508. Reconstruction quality $|\Delta\mathrm{CE}|$ stays within 0.0005 of Fixed on all four cells, so the causal-metric gains do not destabilize the well-aligned dictionary GAE starts from. This is consistent with the geometric picture: GAE's correction is driven by the gap between the explainer's encoded covariance and the covariance of the evaluated activations, and that gap does not disappear simply because the two are drawn from the same nominal distribution. The same mechanism that recovers faithfulness under explicit domain, temporal, and adversarial shift continues to operate, at smaller magnitude, on the residual within-corpus drift that ID held-out evaluation exposes.

H Evaluation on Llama-3.1-8B with LlamaScope

Setup and faithfulness. We evaluate a public LlamaScope $32\times$ Top-K SAE [45] on Llama-3.1-8B under the HaluEval protocol used in the main experiments. The hidden dimension is $d = 4{,}096$, the dictionary size is $k = 131{,}072$, and the adaptation rank is $r = 64$. GAE uses $N = 2{,}048$ unlabeled OOD hidden-state activation vectors while keeping the model and dictionary encoder frozen. Table 5 compares GAE with Fixed, Finetune, and SAEBoost. GAE achieves the highest nAOPC and nComp and the lowest $|\Delta\mathrm{CE}|$ among the evaluated methods. Relative to Fixed, nAOPC increases from 0.5719 to 0.5972, nComp from 0.7892 to 0.7936, and $|\Delta\mathrm{CE}|$ decreases from 0.0652 to 0.0602. This evaluation supports application to one larger public SAE on HaluEval.

Measured computational cost. Table 12 compares GAE adaptation with OOD fine-tuning on an RTX A6000. The GAE time includes second-moment estimation, eigenspace computation, and the closed-form decoder update for one rank, excluding hyperparameter search. The fine-tuning run uses 199,936 OOD activation tokens, with 781 steps and batch size 256. GAE reduces adaptation time substantially, while both methods use tens of gigabytes of GPU memory.

Table 12: **Computational cost for Llama-3.1-8B with LlamaScope on HaluEval.** Time and GPU memory are approximate measurements on an RTX A6000.

Method	Time	GPU memory
OOD fine-tuning	$\sim$5 h 25 min	$\sim$32 GB
GAE	**$\sim$30 s**	**$\sim$28 GB**

Scaling considerations. Forming the dense $d \times d$ OOD second-moment matrix and computing its eigenspace requires $O(Nd^2 + d^3)$ time and $O(d^2)$ storage for the matrix. The decoder itself contains dk entries, so increasing the dictionary size also increases decoder storage and update cost. GAE avoids gradient updates and optimizer states, but dense eigenspace computation remains a bottleneck as d grows. For rank sweeps, the OOD second moment and eigenvectors can be reused, with each additional rank changing the cutoff and decoder update. Streaming second-moment accumulation and scalable eigensolvers are potential extensions for larger hidden dimensions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's contributions: formalizing the faithfulness gap, proposing GAE, proving theoretical guarantees, and evaluating the method across multiple OOD settings. The scope is also qualified by the evaluated models and shifts, with Llama-3.1-8B evaluated only on HaluEval (Appendix H).

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the conclusion, including the top-r SVD truncation of the ID dictionary, and future directions such as adaptive rank selection and encoder adaptation. Appendix H discusses the cost of dense eigenspace computation as the hidden dimension grows.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper states assumptions for its main theoretical results, including eigengap assumptions for Propositions 1 and 2 and Theorem 1. Complete proofs and derivations are provided in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the model/explainer configurations, OOD settings, baselines, evaluation metrics, GAE hyperparameters, evaluation protocol, and evaluation size needed to reproduce the main experimental comparisons. Appendix D gives the detailed experimental and implementation setup, and the public code repository is available at `https://github.com/MLAI-Yonsei/GAE/` to further support reproduction.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides a public code repository at https://github.com/MLAI-Yonsei/GAE/. The repository README documents the implementation framework, main runner, supported models and datasets, baselines, shared adaptation/evaluation protocol, implemented metrics, CLI entrypoints, output contracts, and reproducibility checks; setting-specific OOD details are documented in separate repository files.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental setup specifies the target models, explainer types, dictionary sizes, ID and OOD datasets, baselines, evaluation metrics, feature budgets, evaluation protocol, and evaluation size. Appendix D provides additional model/explainer configurations, baseline descriptions, GAE implementation details, and GAE hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We report metrics as the mean over 1,000 evaluation prompts per setting, without error bars or confidence intervals for these main comparisons. Table 11 reports paired-bootstrap 95% confidence intervals for the concept-level DLA differences. We use a single training seed (2026) for stochastic baselines due to compute constraints (Appendix D.6).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information about the computer resources in Appendix D.6.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and are not aware of any deviations. The work evaluates mechanistic interpretability methods on existing pretrained language models and existing text datasets/benchmarks, without human-subject experiments, crowdsourcing, or deployment on users.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The current paper is primarily a methodological study of faithful dictionary-based interpretability under distribution shift and does not include a dedicated discussion of both positive and negative societal impacts. Because the method is evaluated as a research tool on existing frozen pretrained language models and existing text datasets/benchmarks, without a user-facing deployment, human-subject study, or release of a new high-risk model or dataset, the paper does not provide an application-specific societal-impact analysis; broader-impact considerations are left for future versions.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new pretrained language model, image generator, scraped dataset, or other high-risk artifact requiring controlled release safeguards. The released asset is an implementation of the proposed explainer adaptation method.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses existing pretrained models, datasets, and benchmarks, and credits the original creators of the main experimental assets through citations. The authors use third-party assets in accordance with their applicable licenses and terms of use, without repackaging or redistributing restricted model checkpoints or datasets. The paper submission is designated under CC-BY 4.0.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The main new asset is the released implementation of GAE, available at `https://github.com/MLAI-Yonsei/GAE/`. The method is documented in the paper through Algorithm 1 and Appendix D, including implementation details and reported hyperparameters; the accompanying repository documentation further provides protocol, execution, output, and reproducibility information.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing, user studies, or other human-subject experiments. All experiments are conducted on pretrained language models and existing text datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human-subject research, crowdsourcing, or participant data collection, so IRB approval or equivalent review is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper explicitly describes the use of pretrained language models, including GPT-2 Small and Pythia-1.4B, as frozen target models for evaluating dictionary-based explainers. LLMs are not used as a non-standard component for data annotation, data generation, method construction, or evaluation beyond this stated model-evaluation setting.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.